

Using reforecasts to improve probabilistic weather forecast guidance

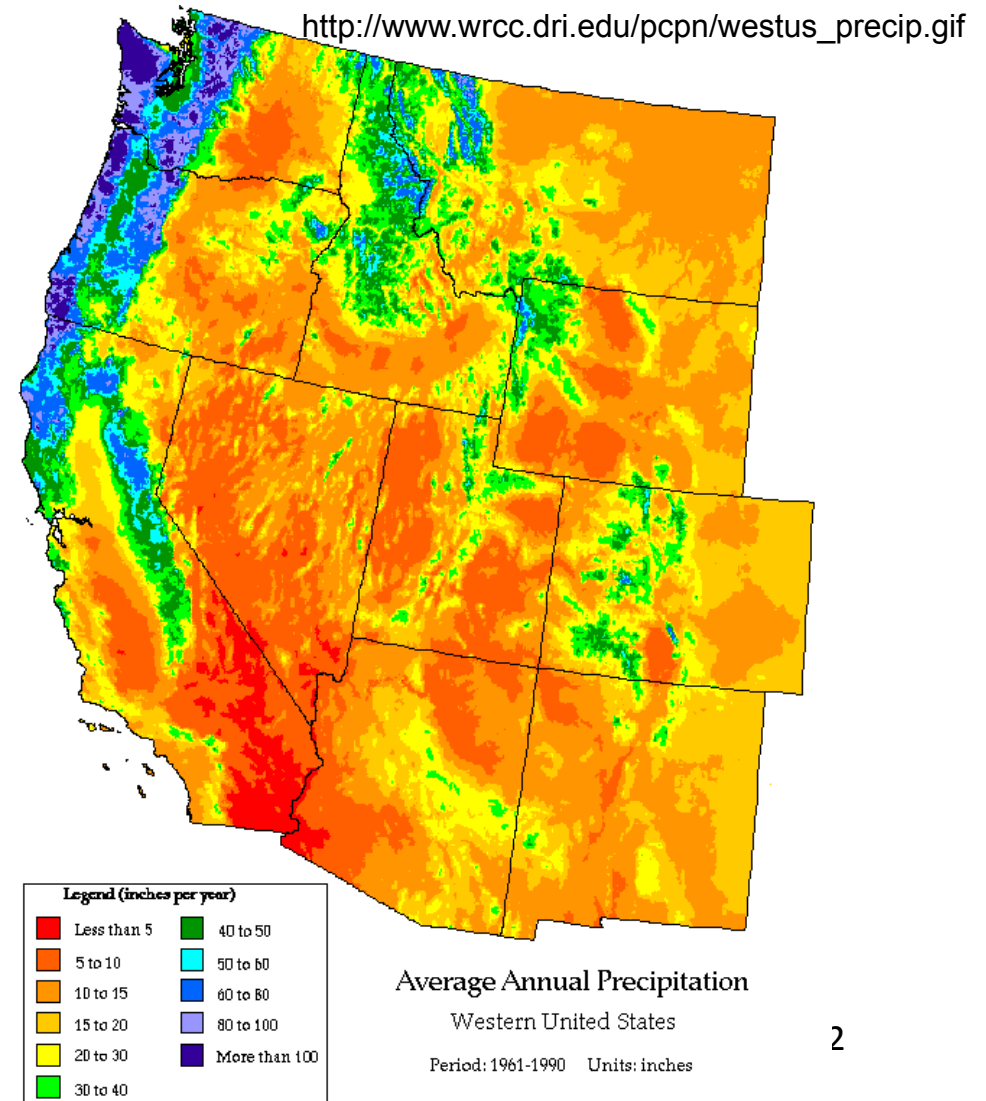
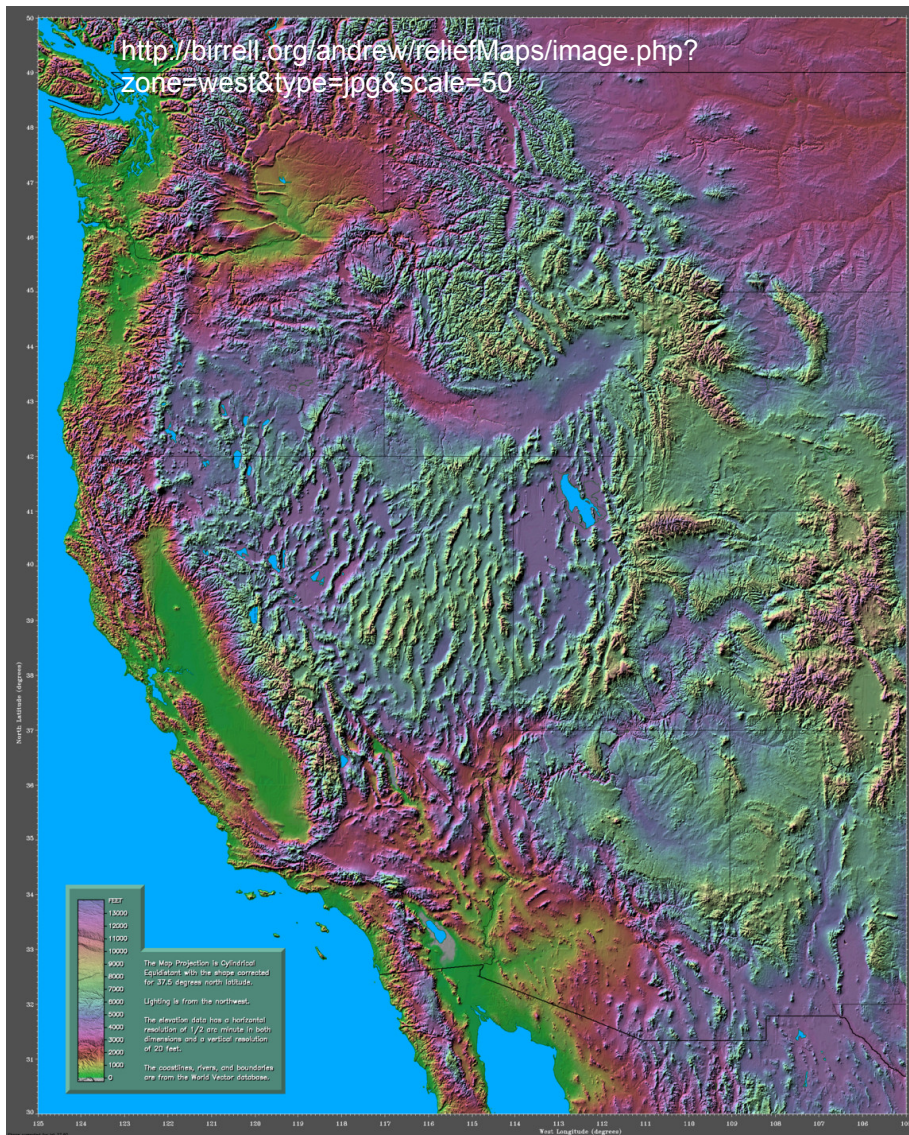
Tom Hamill

NOAA Earth System Research Lab, Boulder, CO USA

tom.hamill@noaa.gov

Terrain and precipitation patterns in western US.

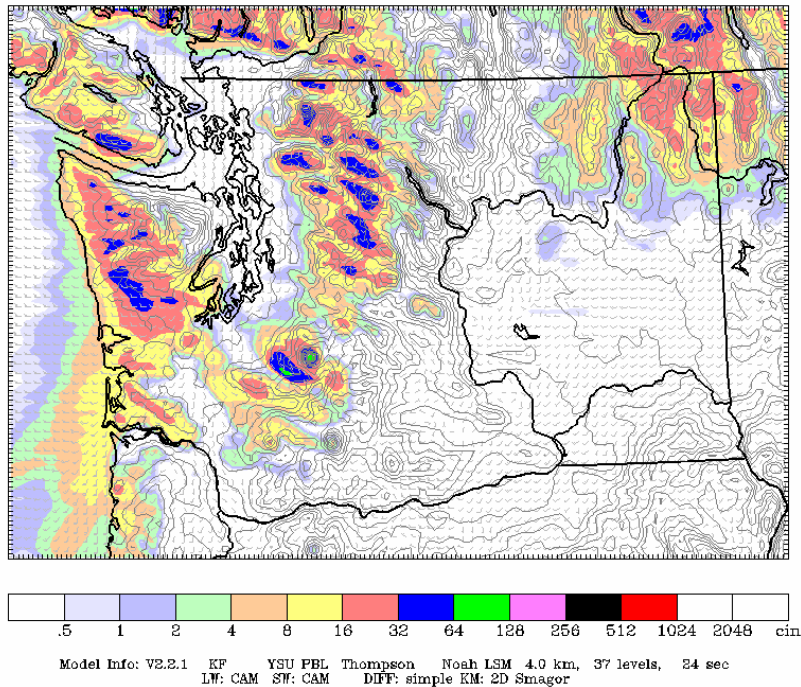
We have mountains, too, and they modulate our weather.



How can we provide spatially detailed weather forecasts in such regions?

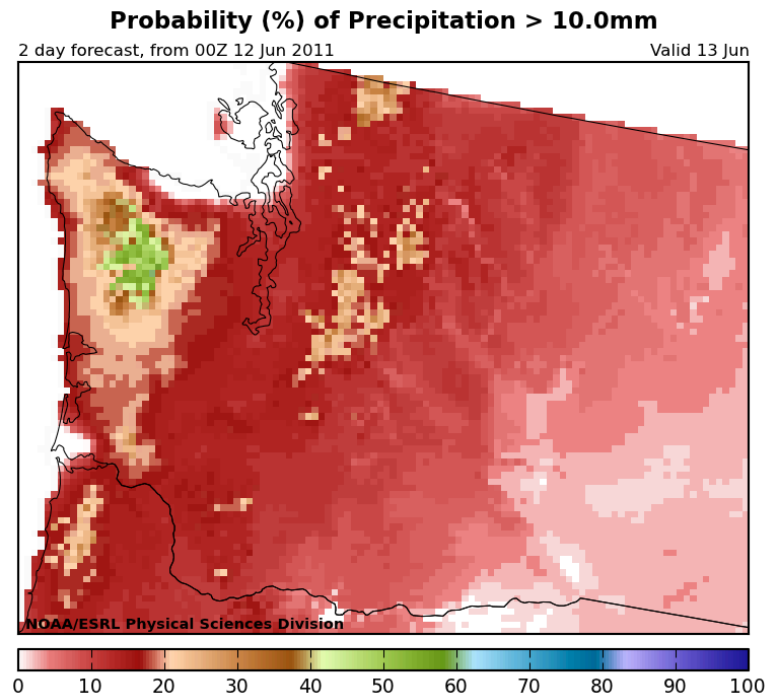
(1) Dynamical downscaling;
high-resolution NWP

UW WRF-GFS 4km Domain
Fcast: 24 h
Valid: 12 UTC Wed 27 Aug 08 (05 PDT Wed 27 Aug 08)
Init: 12 UTC Tue 26 Aug 08
Total Precip in past 3 hrs (.01in)
Wind at 10m (full barb = 10kts)



<http://www.atmos.washington.edu/~cliff/cliff.php>

(2) Statistical downscaling;
relate high-resolution
measurements to
lower-res. model forecast

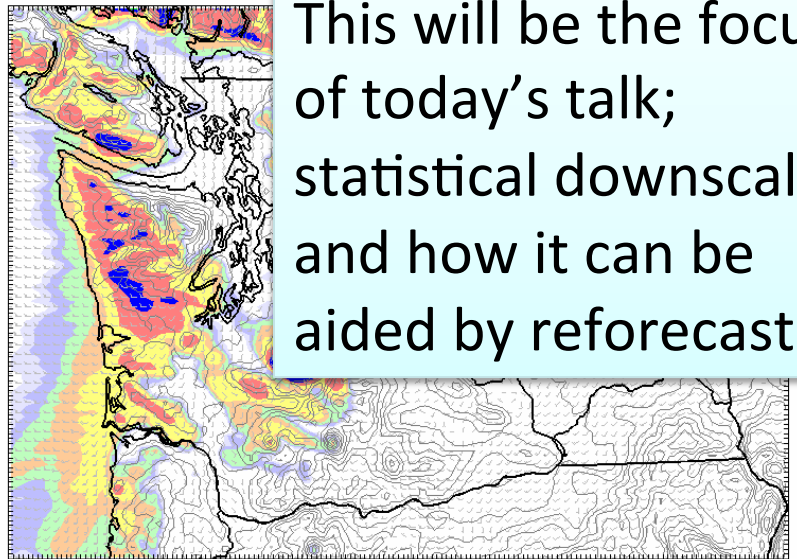


<http://www.esrl.noaa.gov/psd/forecasts/reforecast/narr/>

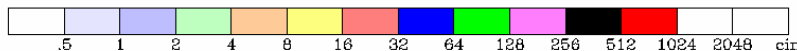
How can we provide spatially detailed weather forecasts in such regions?

(1) Dynamical downscaling;
high-resolution NWP

UW WRF-GFS 4km Domain
Fcast: 24 h
Valid: 12 UTC Wed 27 Aug 08 (05 PDT Wed 27 Aug 08)
Init: 12 UTC Tue 26 Aug 08
Total Precip in past 3 hrs (.01in)
Wind at 10m (full barb = 10kts)



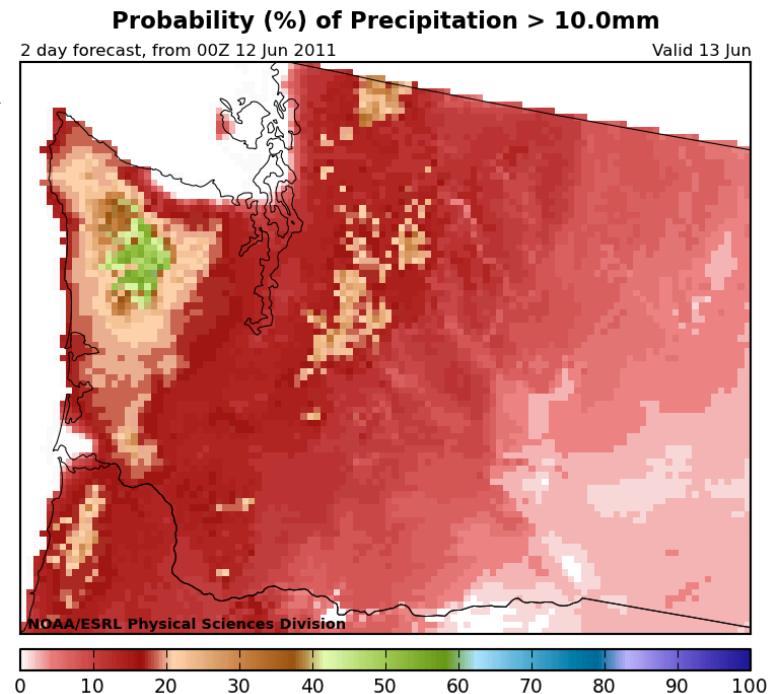
This will be the focus
of today's talk;
statistical downscaling
and how it can be
aided by reforecasts.



Model Info: V2.2.1 KF YSU PBL Thompson Noah LSM 4.0 km, 37 levels, 24 sec
LW: CAM SW: CAM DIFF: simple KM: 2D Smagor

<http://www.atmos.washington.edu/~cliff/cliff.php>

(2) Statistical downscaling;
relate high-resolution
measurements to
lower-res. model forecast



<http://www.esrl.noaa.gov/psd/forecasts/reforecast/narr/>

Definition

- **Reforecasts** are *retrospective* numerical *forecasts*, ideally conducted using the same model, the same data assimilation system that is used operationally. Also called “hindcasts”
- We prefer “reforecast” to emphasize connection with reanalyses and having quality initial conditions.

Outline

- Why generate reforecasts? Advantages and disadvantages of relying on statistical downscaling to produce forecasts.
- What reforecast data sets are/will be available for NWP?
 - details on our upcoming reforecasts.

Why generate and use
reforecasts?

Why not?

What can post-processing with reforecasts do that other NWP techniques cannot?

- Provide extra “resolution,” but via statistical downscaling.
 - Also: compensate for systematic model biases, thereby increase reliability, increase forecast skill.
 - High-resolution models may have lots of detail but are not free from bias!
- Provide sufficient samples to quantify forecast errors for particular locations, hydrologic basins.
- Provide context on how unusual today’s forecast event is, relative to other *forecast* events.

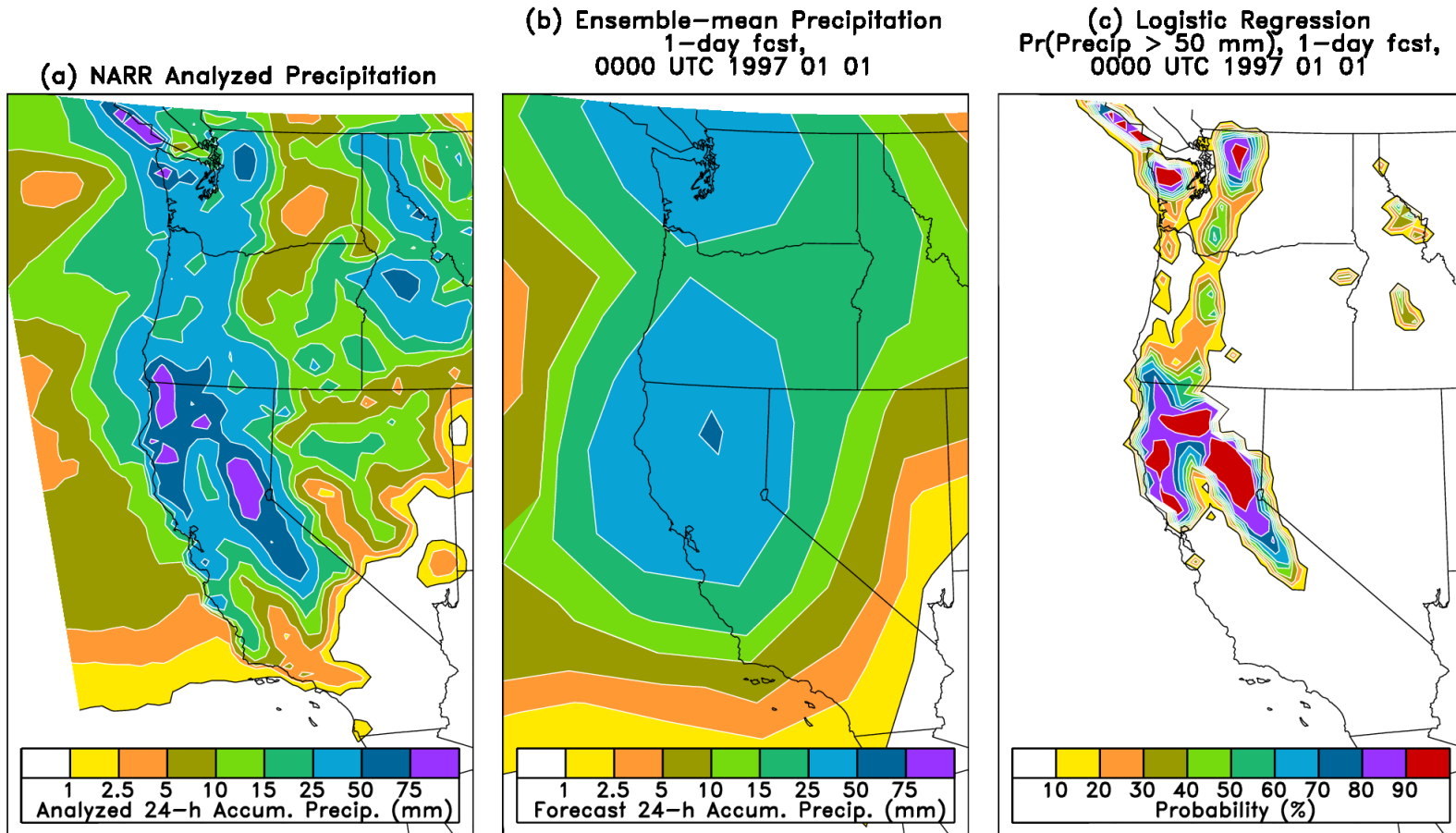
Disadvantages of post-processing with reforecasts

- Right answer perhaps, but for wrong reason? We prefer to directly improve the model in physically realistic ways.
 - Also, some errors are too complex to adjust via post-processing; for these, there is no substitute for improving the model.
- Additional computational and infrastructural burden to compute reforecasts and reanalyses, compile observation time series.
 - ECMWF's (relatively sparse) weekly 5-member reforecast * 20 years = 100 extra members / week to compute.
 - Generally greater benefit the more years, more days, more members in reforecast, but proportionally more expensive.
 - Without high-quality, long observation time series, many of the benefits of reforecasts + statistical post-processing are lost.
 - Need to keep computing reforecasts with current model version, else improvements are temporary.
- If real climate or model-error statistics change significantly during reforecast period, decreased accuracy of post-processed estimates

Post-processing and reforecast advantages

(all results using T62 GFS reforecast
unless otherwise noted)

Statistical downscaling example



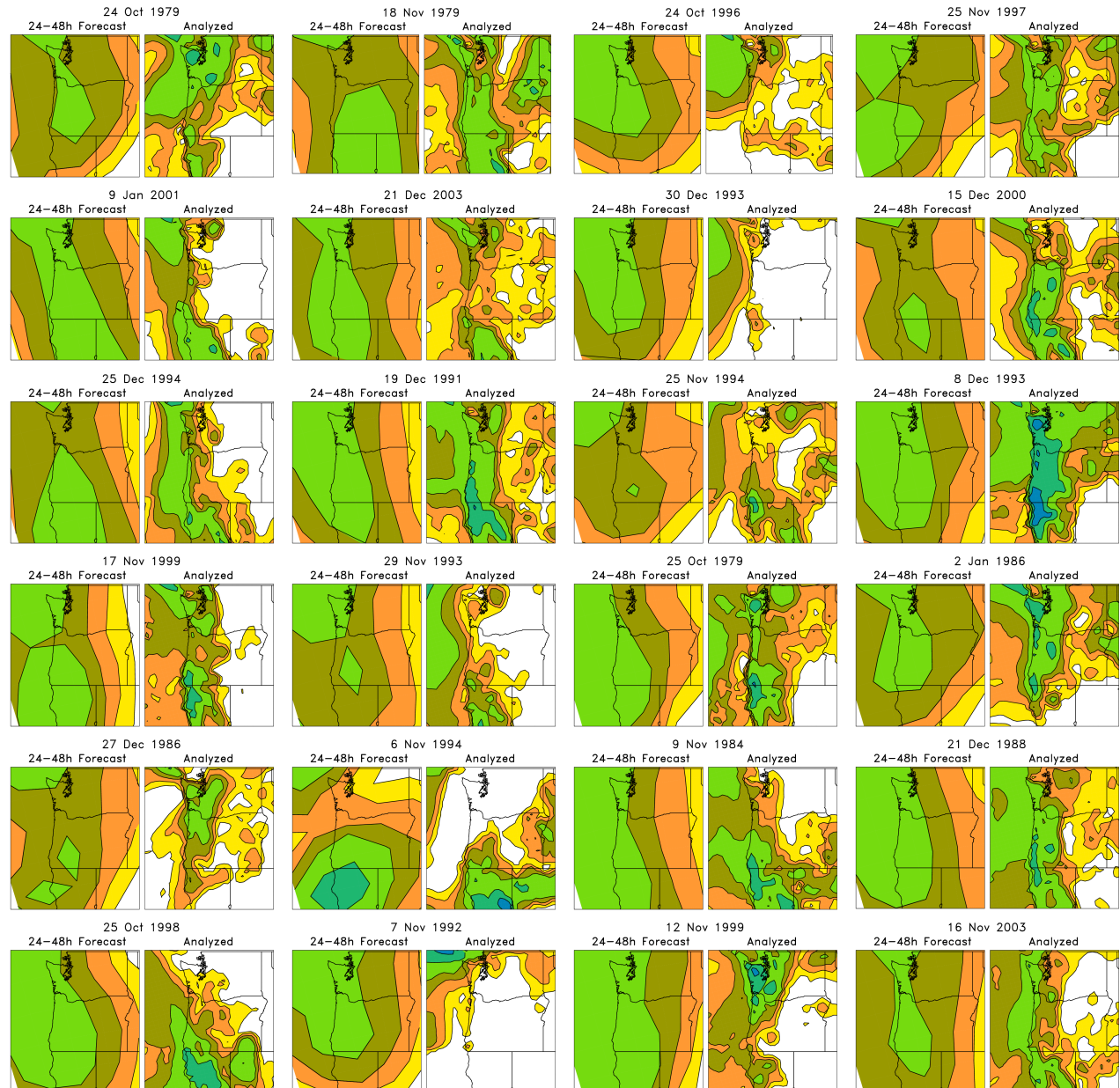
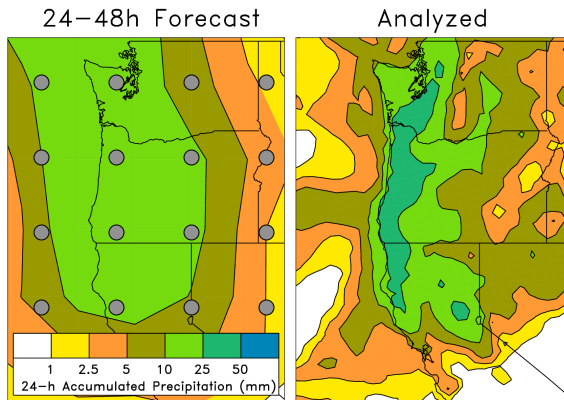
Coarse resolution model forecasts the large-scale precipitation anomaly realistically, but without small-scale detail. Statistical downscaling fills in this detail, relating past forecasts and observations to correct the real-time forecast. Methods of statistical downscaling discussed later.

Analog technique for statistical downscaling

Today's ens. mean
forecast + a posteriori
analyzed precip.



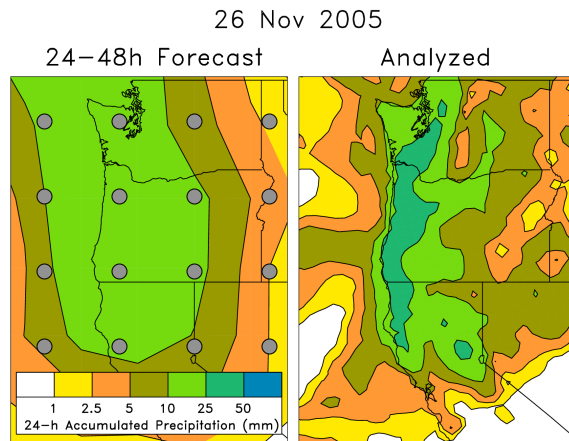
26 Nov 2005



On the left are old forecasts
similar to today's ensemble-
mean forecast. For making
probabilistic forecasts,
form an ensemble from
the accompanying
analyzed weather on the
right-hand side.

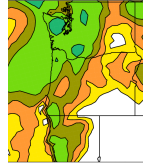


Analog technique for statistical downscaling

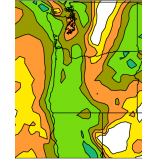


Form an
ensemble from
these

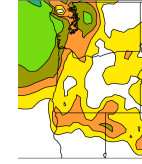
24 Oct 1979
Analyzed



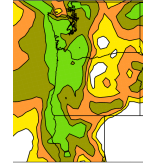
18 Nov 1979
Analyzed



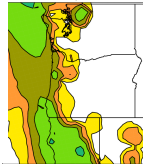
24 Oct 1996
Analyzed



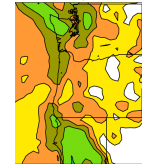
25 Nov 1997
Analyzed



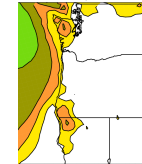
9 Jan 2001
Analyzed



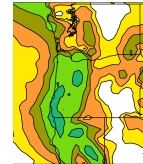
21 Dec 2003
Analyzed



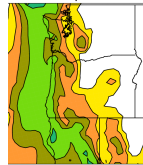
30 Dec 1993
Analyzed



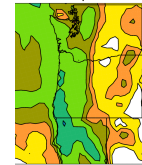
15 Dec 2000
Analyzed



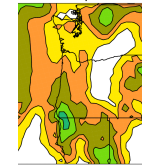
25 Dec 1994
Analyzed



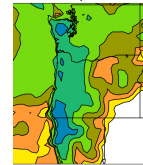
19 Dec 1991
Analyzed



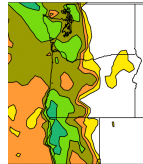
25 Nov 1994
Analyzed



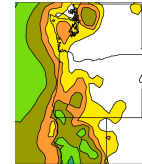
8 Dec 1993
Analyzed



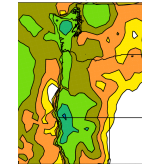
17 Nov 1999
Analyzed



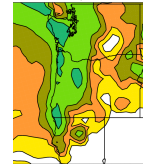
29 Nov 1993
Analyzed



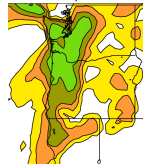
25 Oct 1979
Analyzed



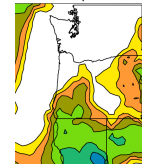
2 Jan 1986
Analyzed



27 Dec 1986
Analyzed



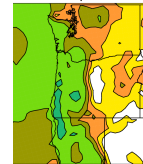
6 Nov 1994
Analyzed



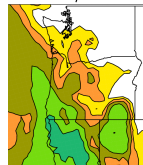
9 Nov 1984
Analyzed



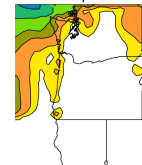
21 Dec 1988
Analyzed



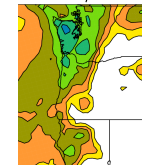
25 Oct 1998
Analyzed



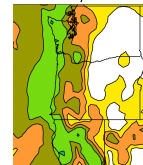
7 Nov 1992
Analyzed



12 Nov 1999
Analyzed



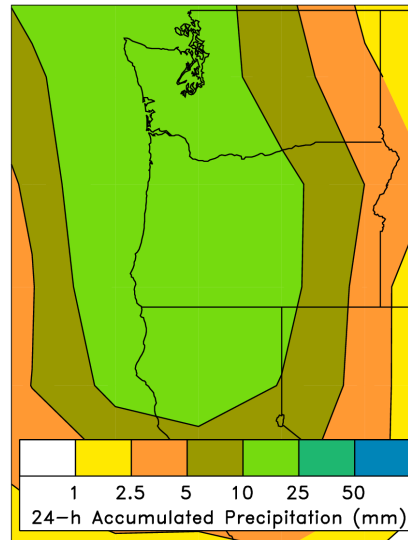
16 Nov 2003
Analyzed



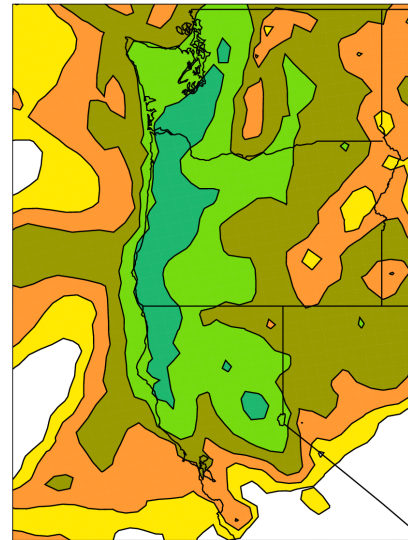
Downscaled analog probability forecasts

26 Nov 2005

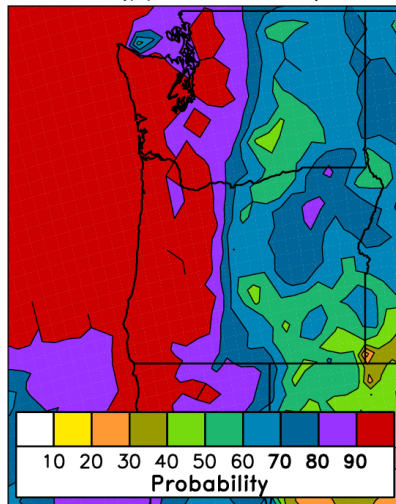
24–48h Forecast



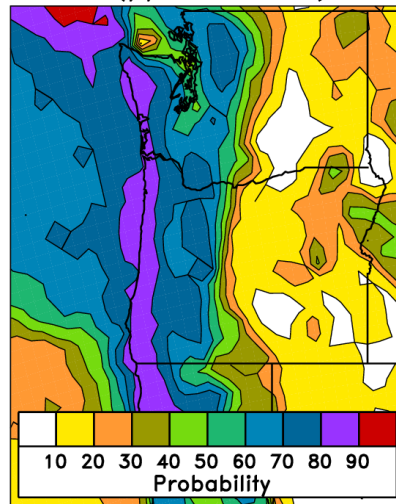
Analyzed



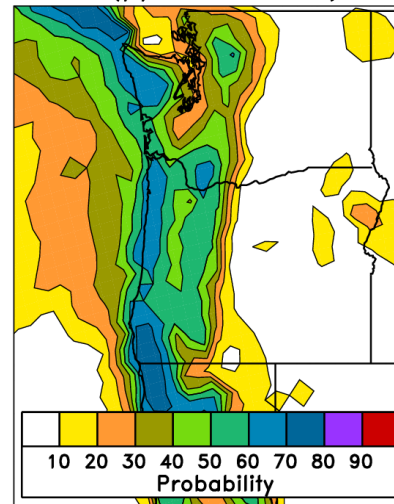
P (ppn > 1 mm)



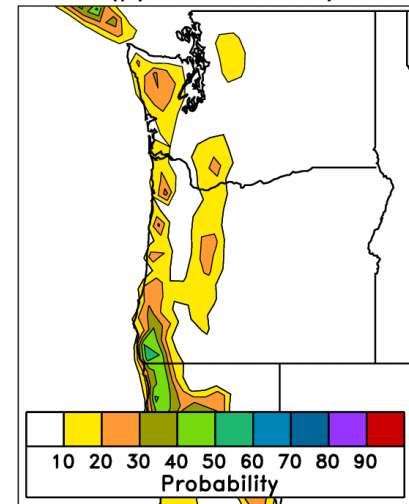
P (ppn > 5 mm)

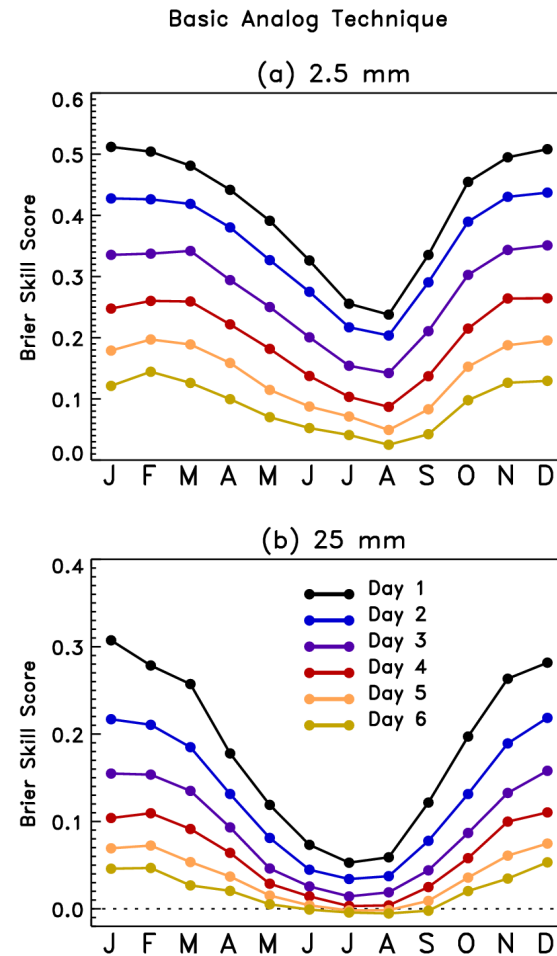
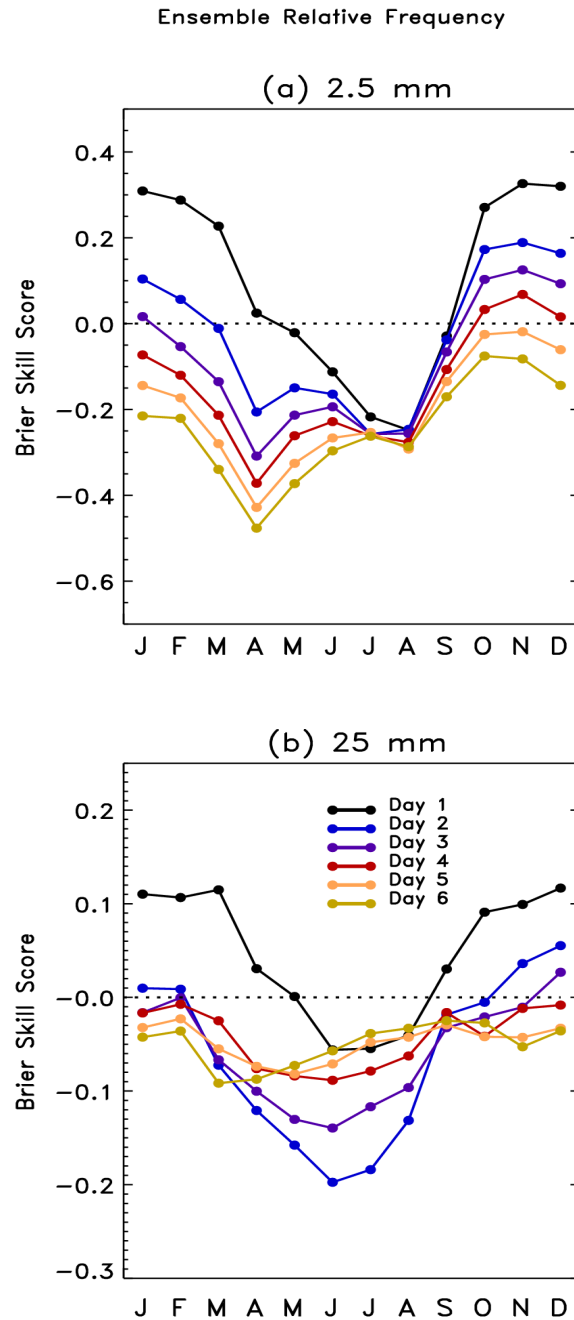


P (ppn > 10 mm)



P(ppn > 25 mm)

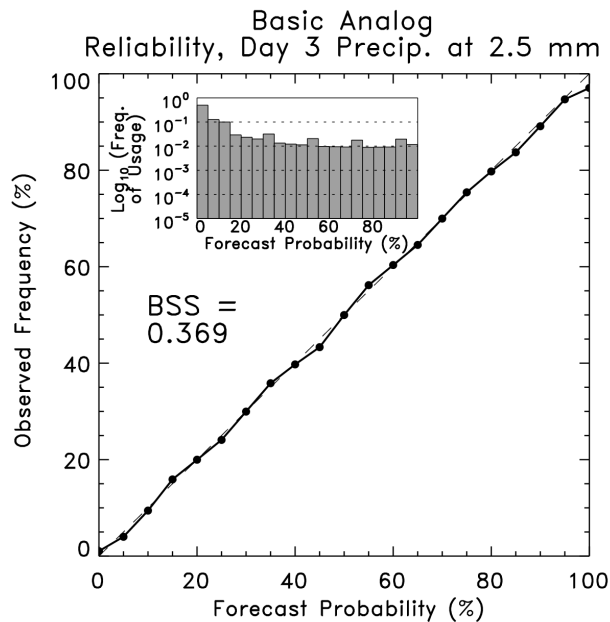




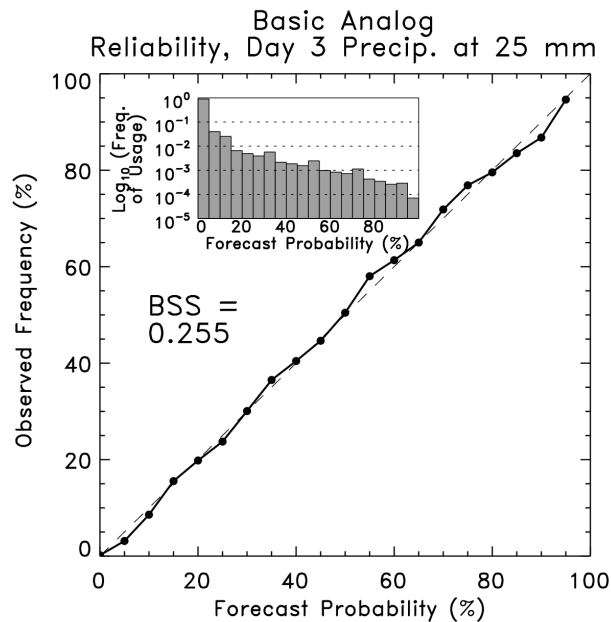
Verified over 25 years of forecasts;
skill scores use conventional
method of calculation which may
overestimate skill
(Hamill and Juras, QJRMS, Oct 2006).

Reliability, analog day +3 forecast

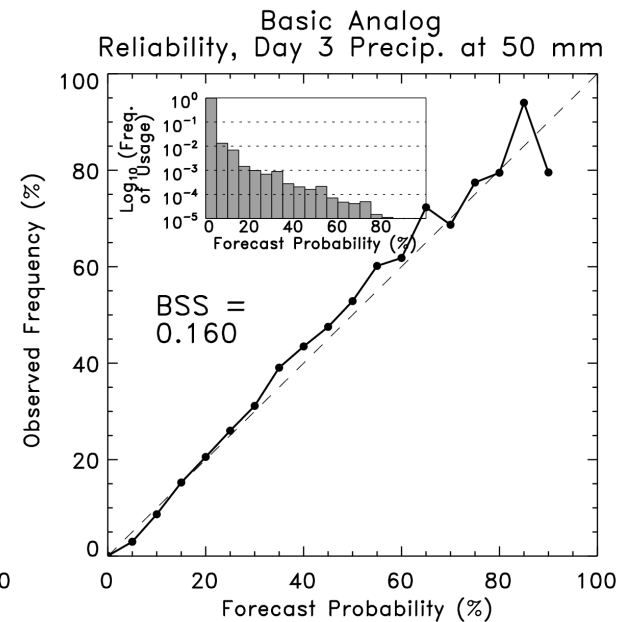
2.5 mm



25 mm



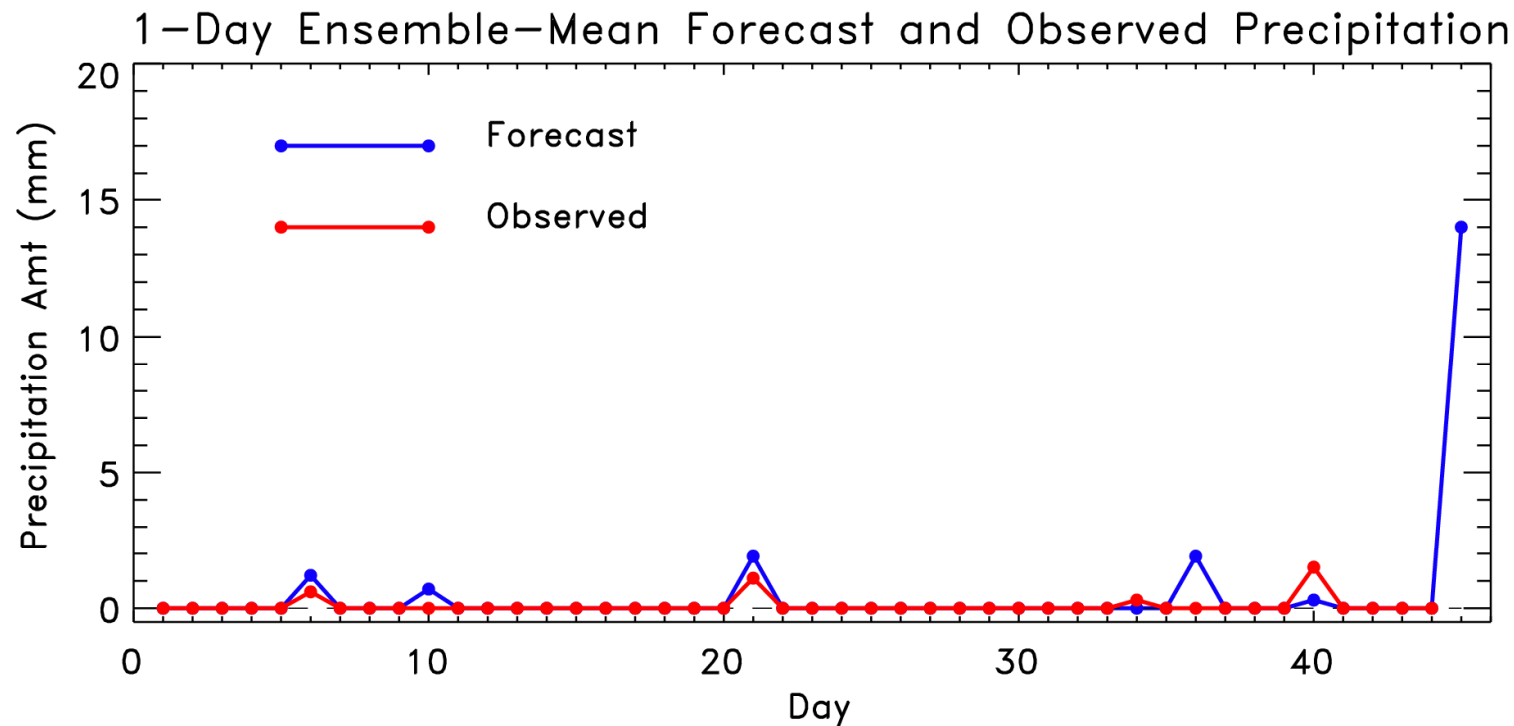
50 mm



Can post-process to achieve reliable, skillful PQPFs

Post-processing of heavy precipitation and other rare events: importance of sample size

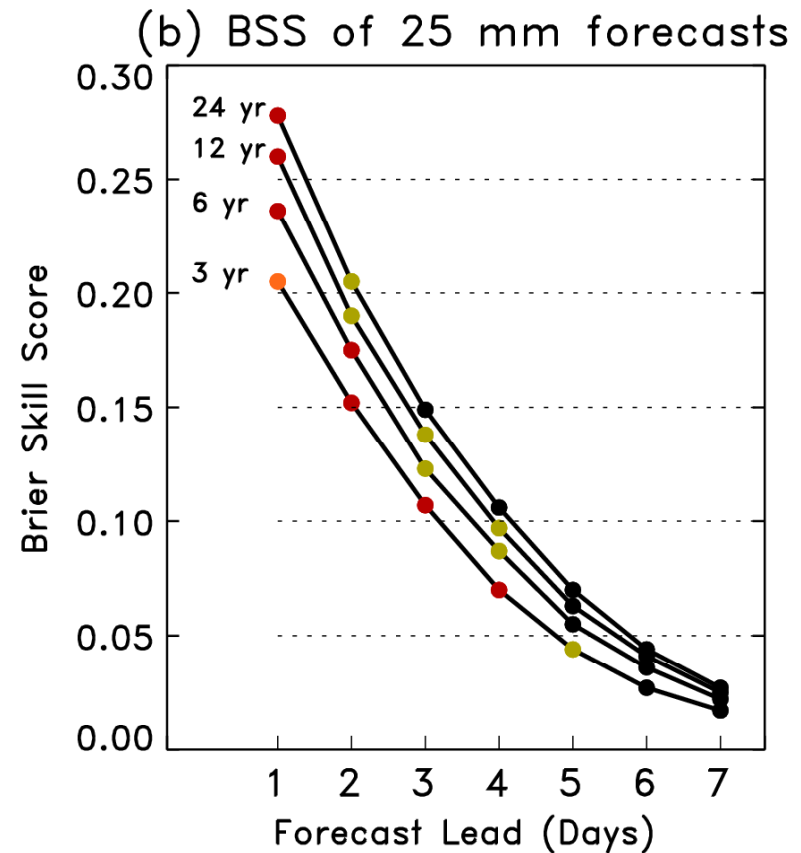
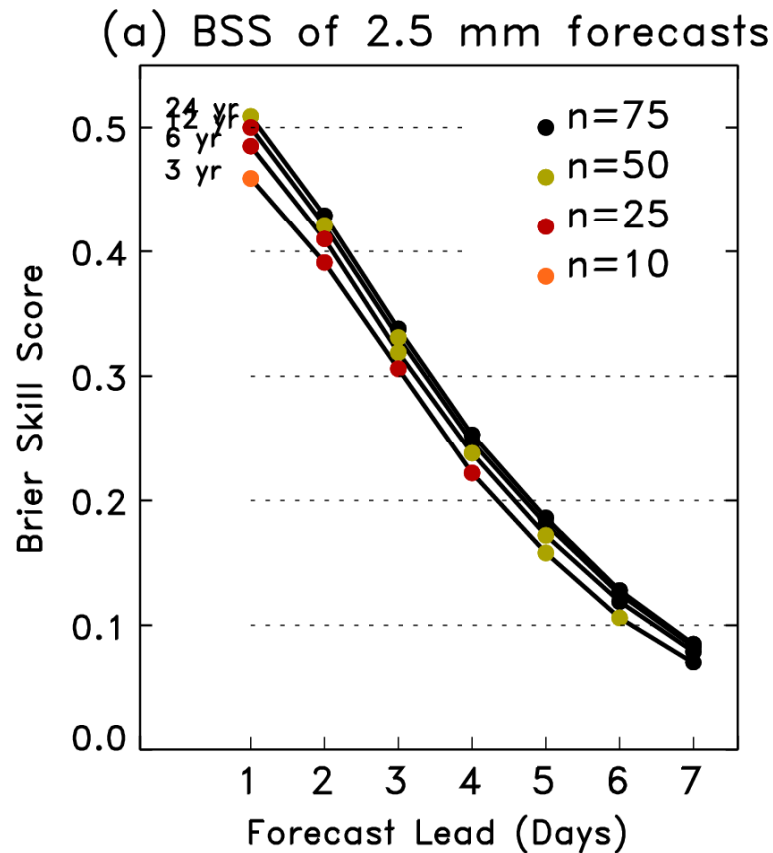
Effective calibration is aided by having old forecast cases that were similar to today's forecast. Then the difference between the observed and forecast on those days can be used to calibrate today's forecast.



In an example like this, the last 45 days of forecasts weren't very useful.

Effect of training sample size:

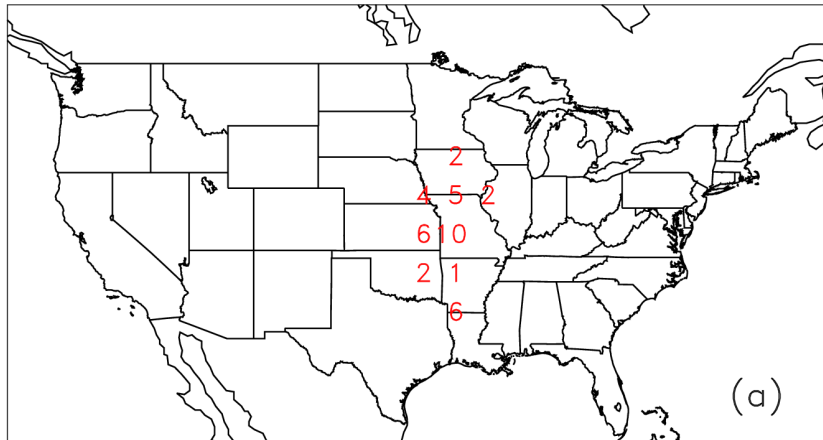
analog technique sensitivity to training sample size



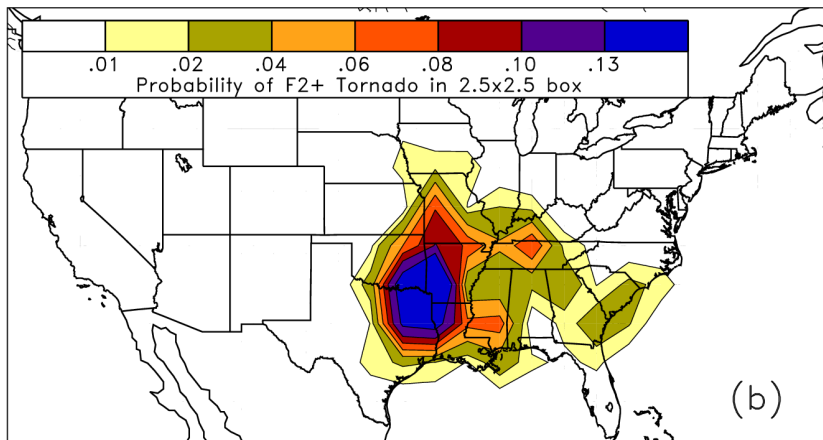
colors of dots indicate which size analog ensemble provided the largest amount of skill.

More exotic post-processing application: tornado probabilities

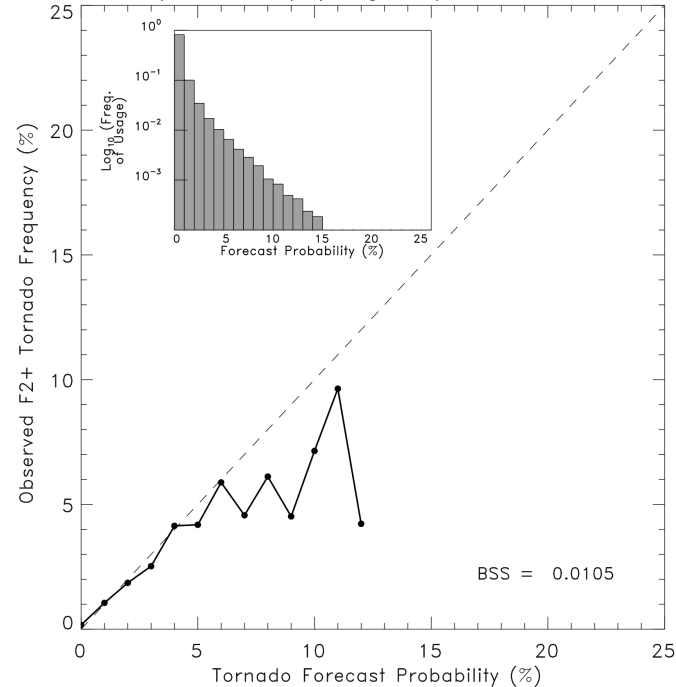
Observed F2+ Tornado Counts in 12-hour Window
Centered on 0000 UTC 27 Apr 1991



Tornado Probabilities for
01-day Forecast from 26 Apr 1991



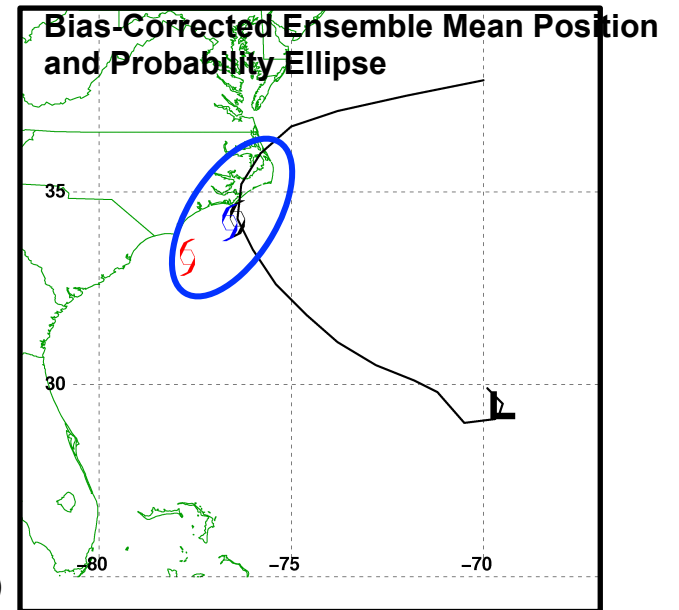
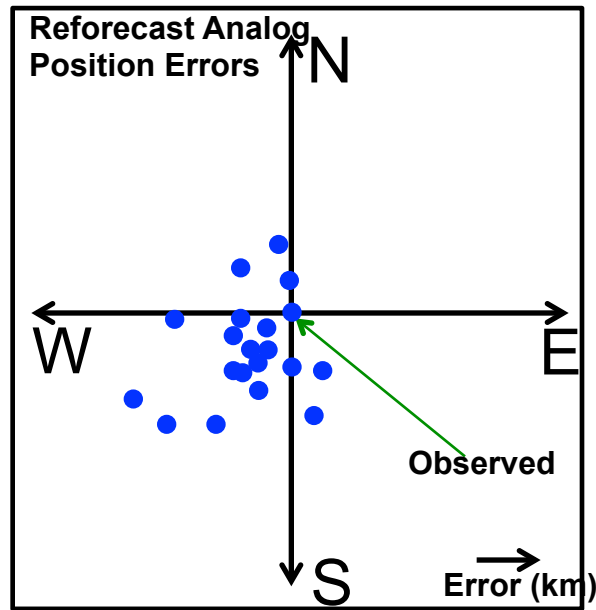
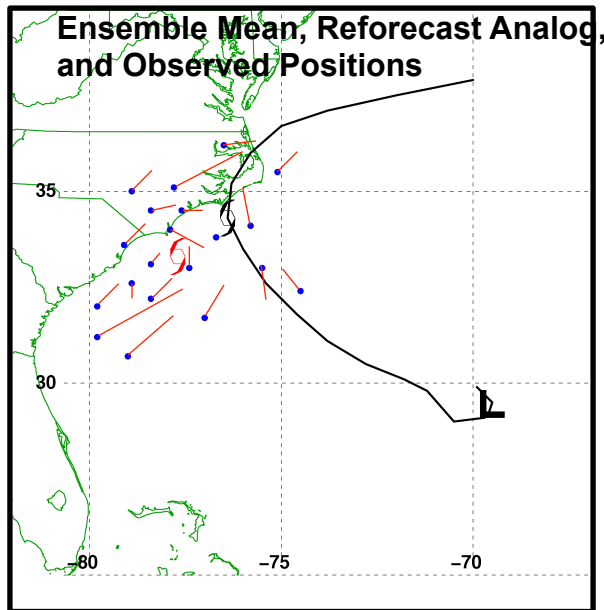
Reliability of 1-Day (Weighted) Tornado Forecasts



We used CAPE, vertical wind shear and analog approach to find dates of similar old cases, then estimated tornado probabilities from occurrence on those dates.

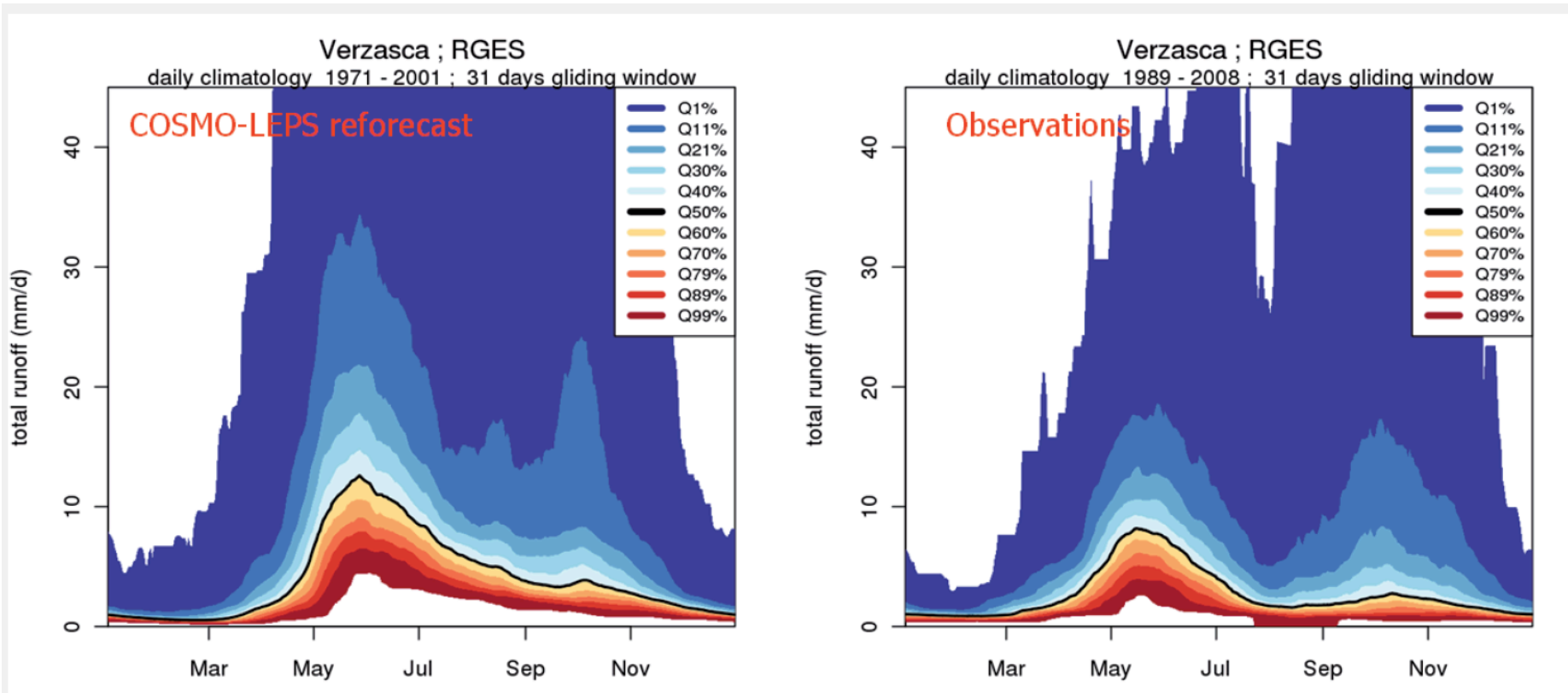
Hurricane track corrections?

72-h Forecast Verifying 1200 UTC 9 September



Another use of reforecasts:

assessing usefulness of model output for hydrologic forecasting

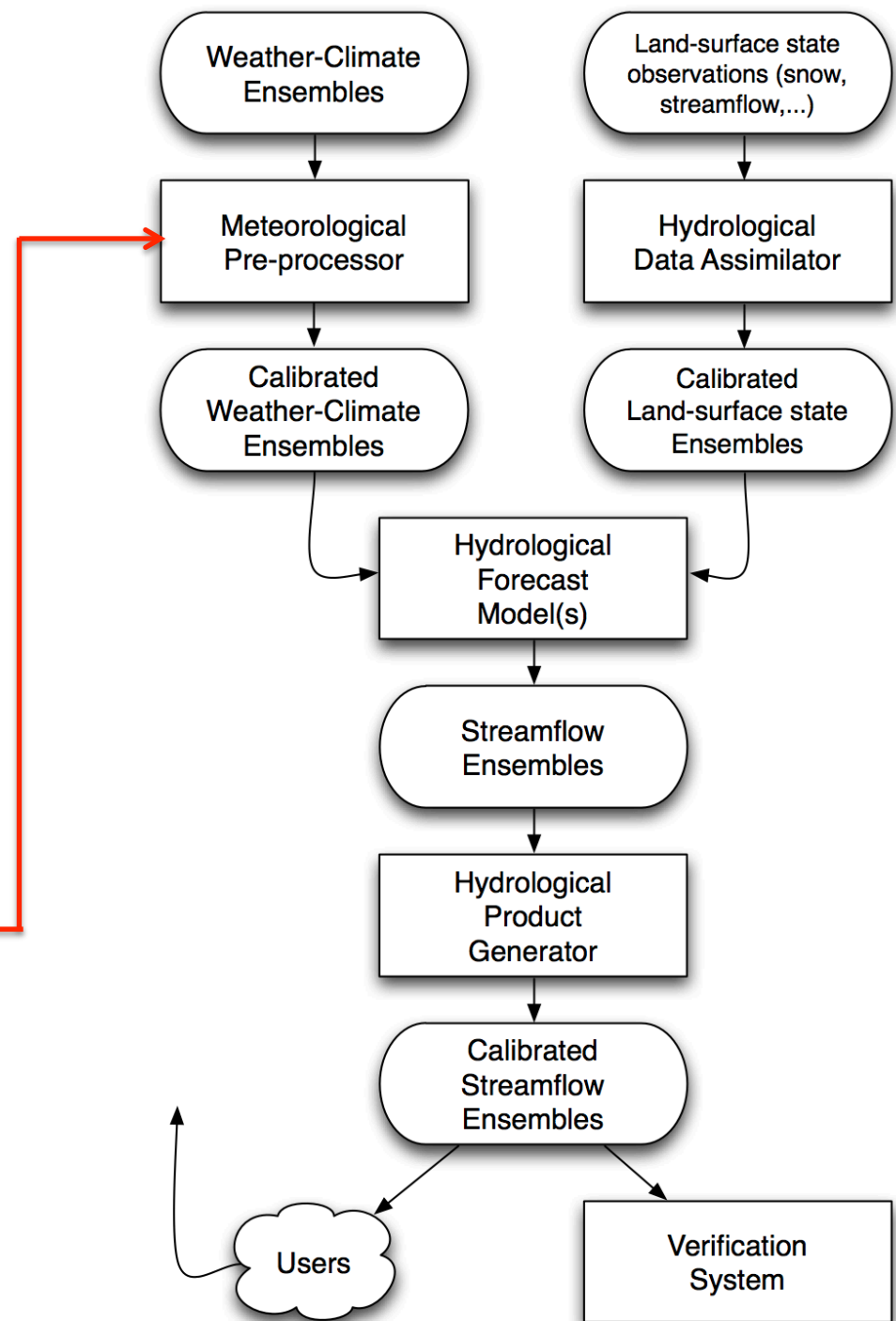


TopLeft: Discharge Climatology Quantiles (30 day gliding mean) for the Verzasca basins obtained forcing the hydrological model PREVAH with COSMO-LEPS reforecasts (1971-2000). TopRight: Observed daily discharge climatology (1989-2008)

Hydrologic Ensemble Prediction Experiment

Note that hydrologists envision a step to make sure that ensemble inputs to their hydrologic system are as reliable and sharp as possible. Reforecasts may be needed for this step.

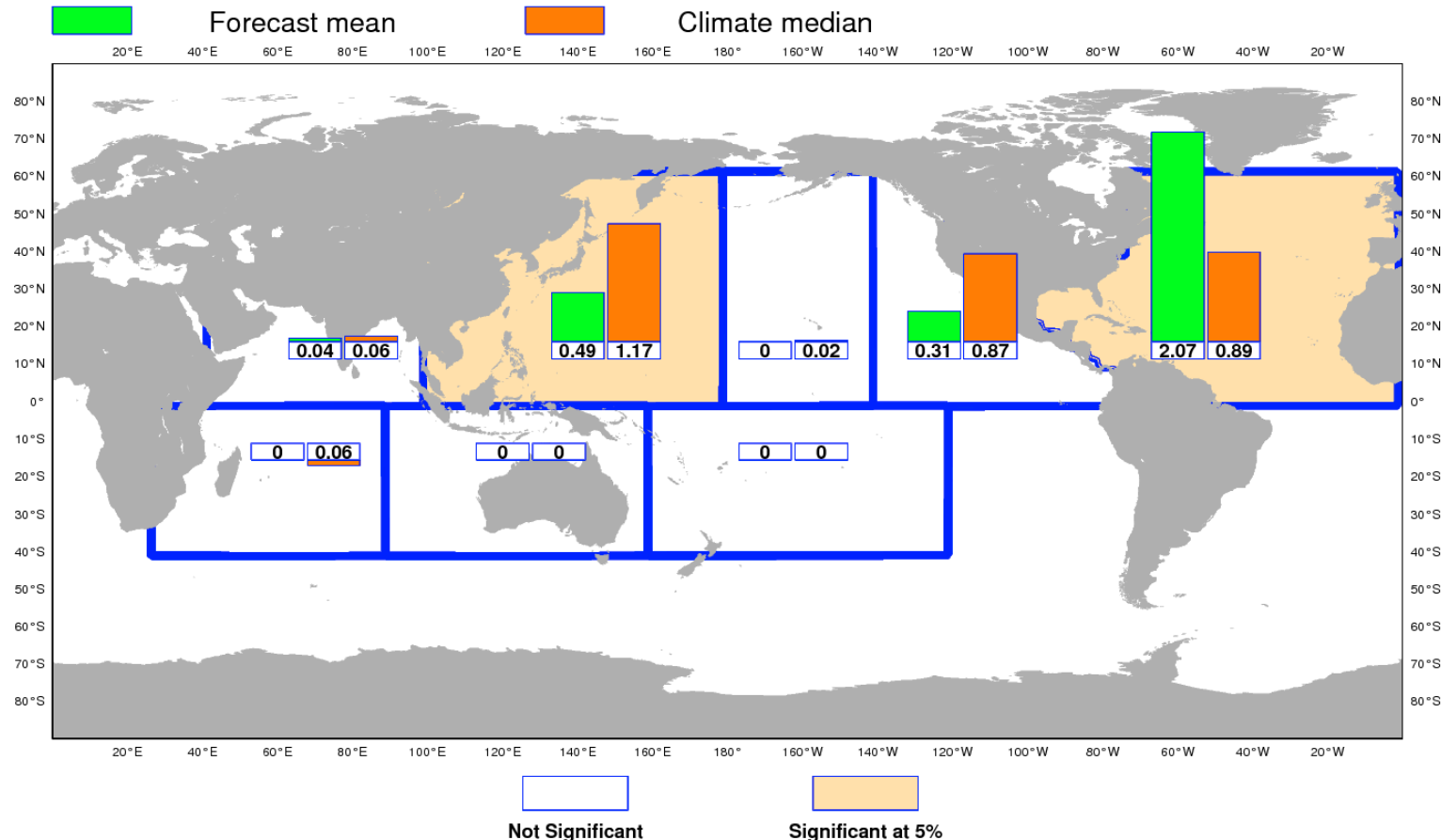
from Schaake et al. 2007 *BAMS* article



Reforecast use: tropical cyclogenesis

ECMWF Monthly Forecast
Tropical Storm Frequency
Forecast start reference is 26/08/2010
Ensemble size = 51, climate size = 90

DAY 12-18
06/09-12/09/2010
Climate = 1992-2009



Many forecast models over-forecast tropical cyclogenesis. This ECMWF product uses TCgenesis from reforecasts to provide some calibration for possible biases.

Ref: D. Richardson, personal communication, ECMWF.

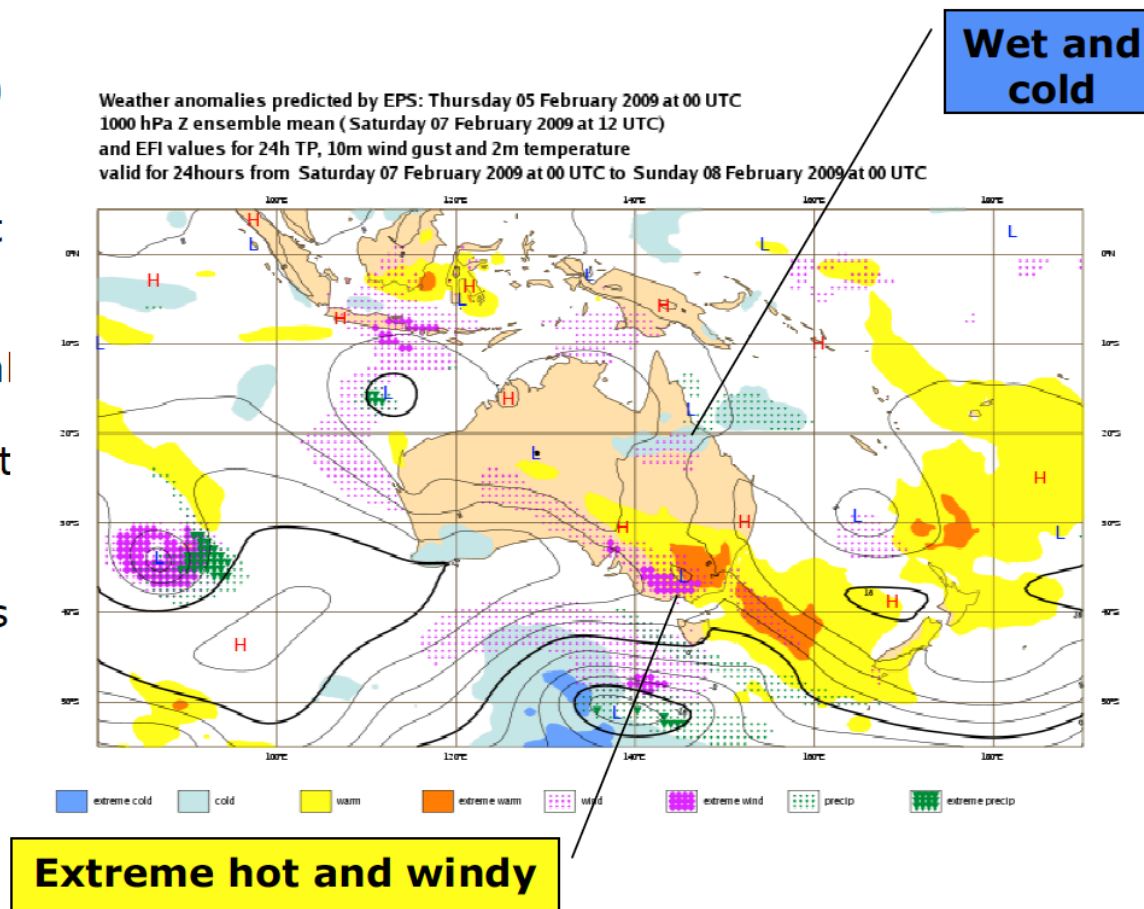
Another reforecast use: facilitates quantitatively assessing how unusual an event is (EFI)



EPS I-EFI 05@00+48/72h vt 07@00-08@00

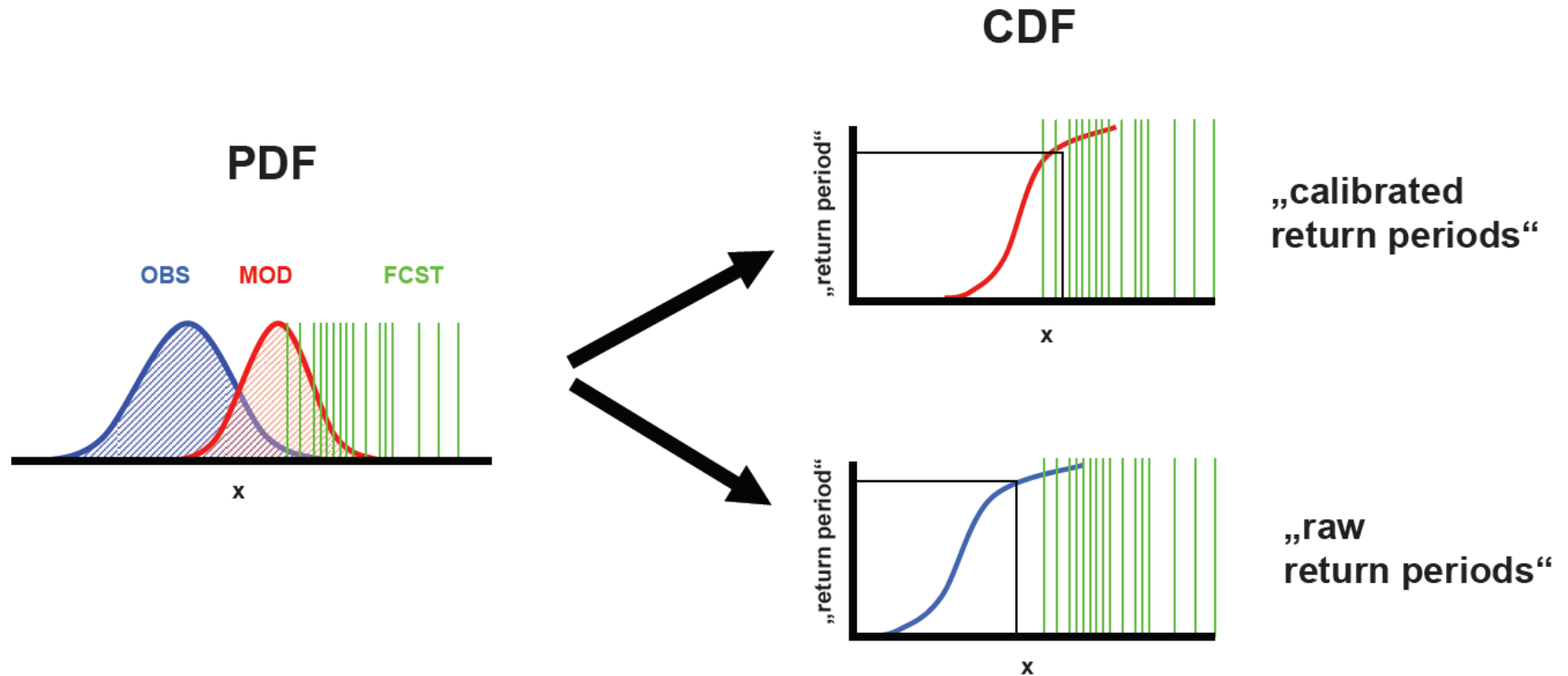
The forthcoming Interactive EFI (I-EFI) can be used to identify areas where the ensemble forecast distribution is significantly different from the climatological distribution, and visualize the grid point distributions.

This plot shows the I-EFI +48/72h forecasts issued on 5@00UTC and valid between 7@00UTC and 8@00UTC.





Calibration strategy



Quantiles w.r.t. observations are not reliable
Quantiles w.r.t model climatology are reliable

Extreme Forecast Index

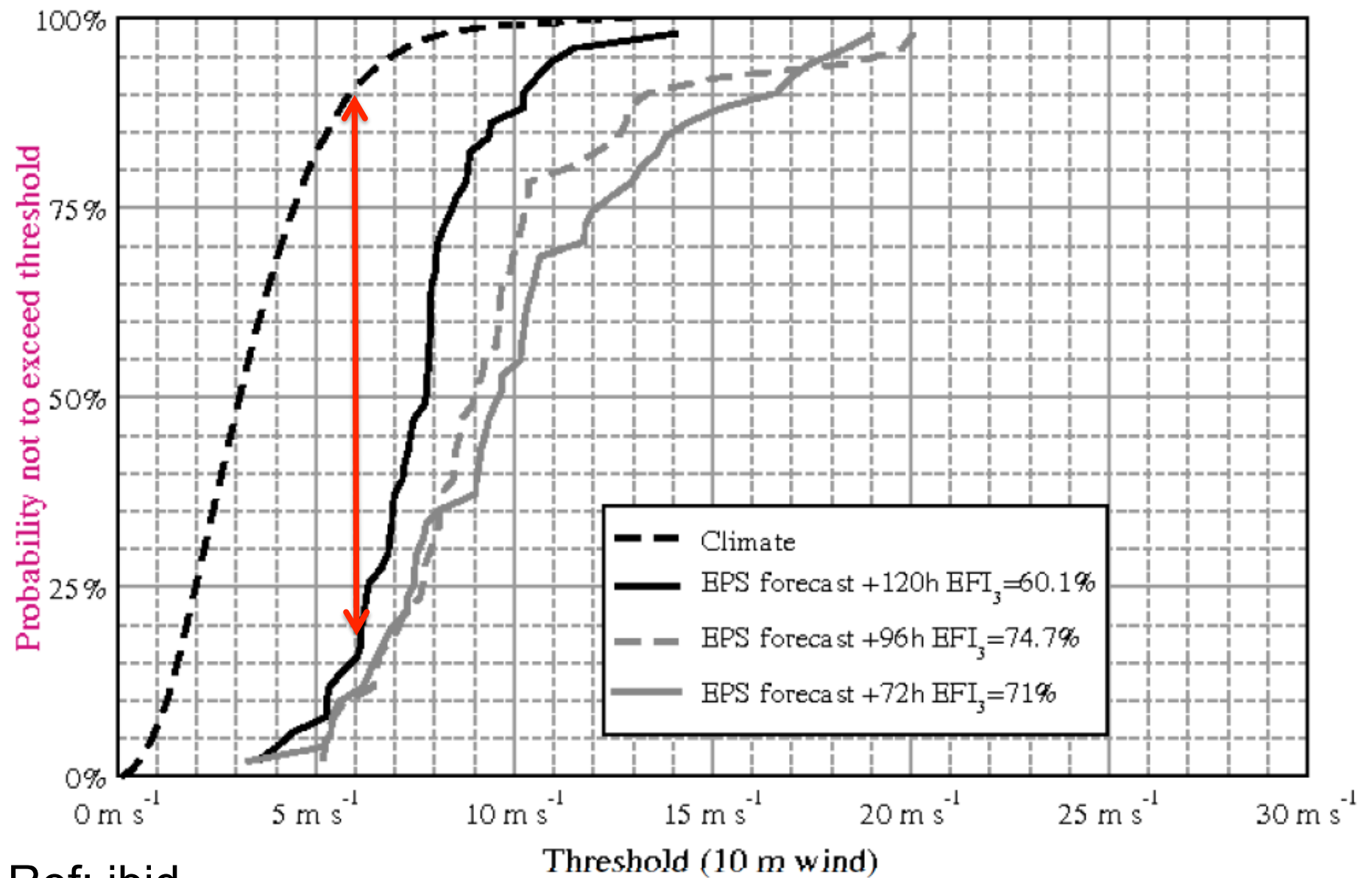
(needs accurate *forecast* climatology,
such as provided by reforecasts)

$$EFI = \frac{2}{\pi} \int_0^1 \frac{p - F_f(p)}{\sqrt{p(1-p)}} dp$$

p is the percentile of the cumulative distribution estimated from the ensemble; $F_f(p)$ is how the p -percentile of the climate record ranks in the EPS (0 the minimum, 1 the maximum). This “Anderson-Darling” version introduces a weighted statistic that gives more power in the tails of the distribution. $2/\pi$ is normalization factor to keep $-1 \leq EFI \leq 1$.

From: LaLaurette, QJRMS, 2003, and http://www.ecmwf.int/products/forecasts/efi_guide.pdf

EFI



Ref: ibid

Reforecast / calibration disadvantages

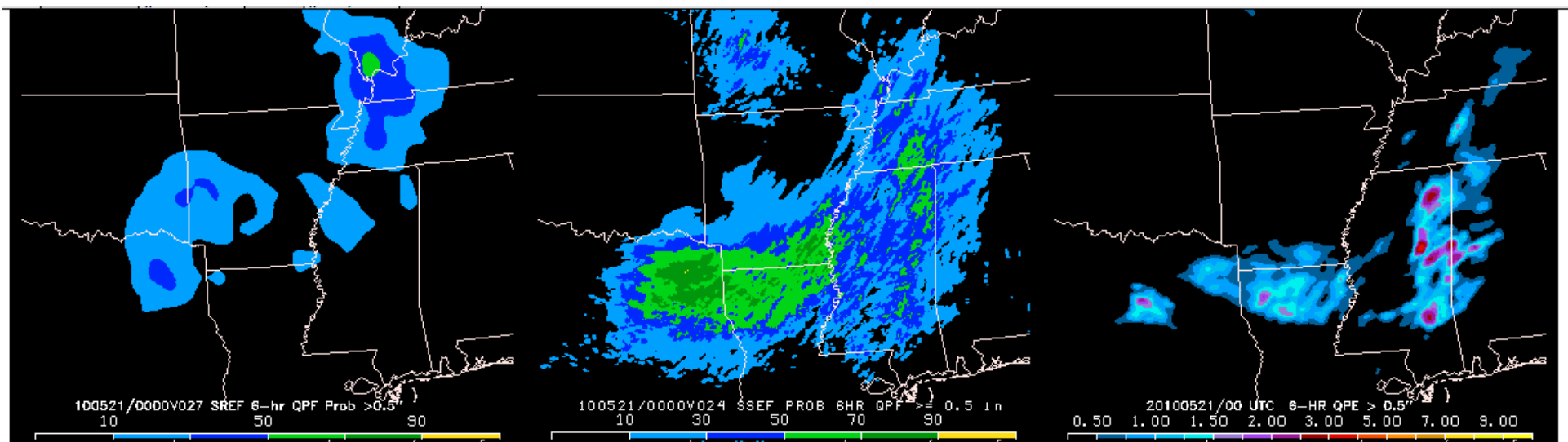
Advantage of *dynamical* downscaling: forecasts of some processes very bad without high resolution, explicit convection

An example from NSSL-SPC Hazardous Weather Test Bed, forecast initialized 20 May 2010
<http://tinyurl.com/2ftbvgs>

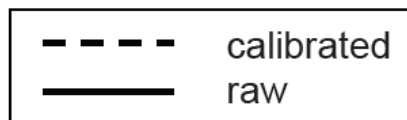
30-km SREF P > 0.5"

4-km SSEF P > 0.5 "

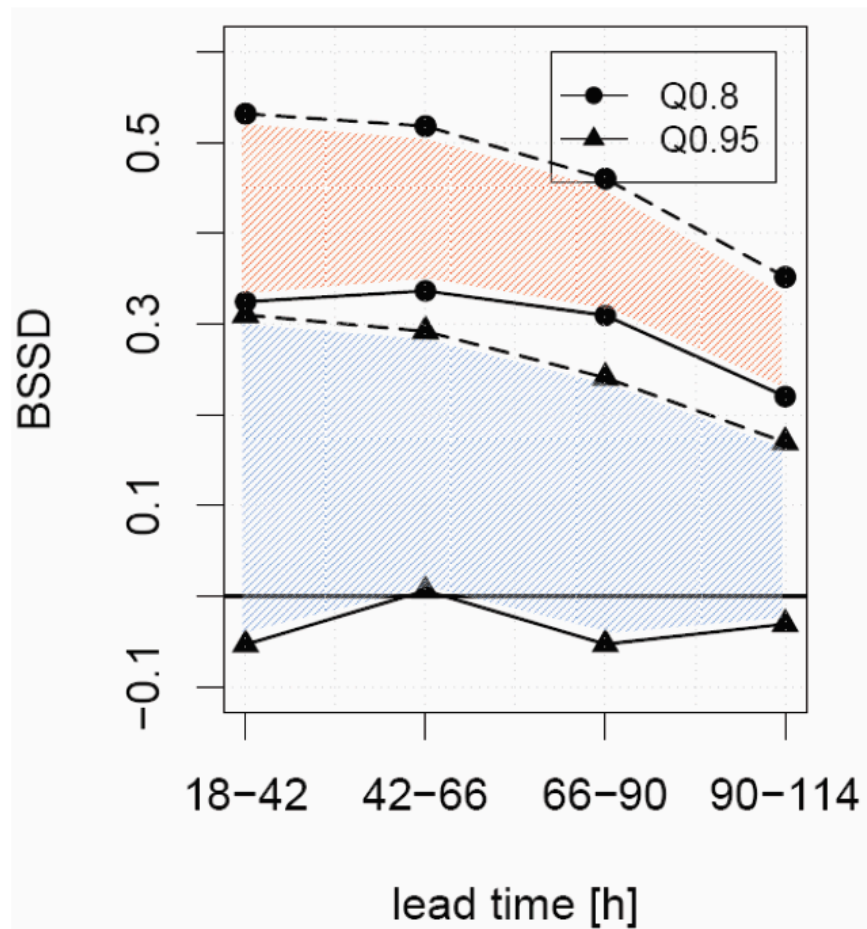
Verification



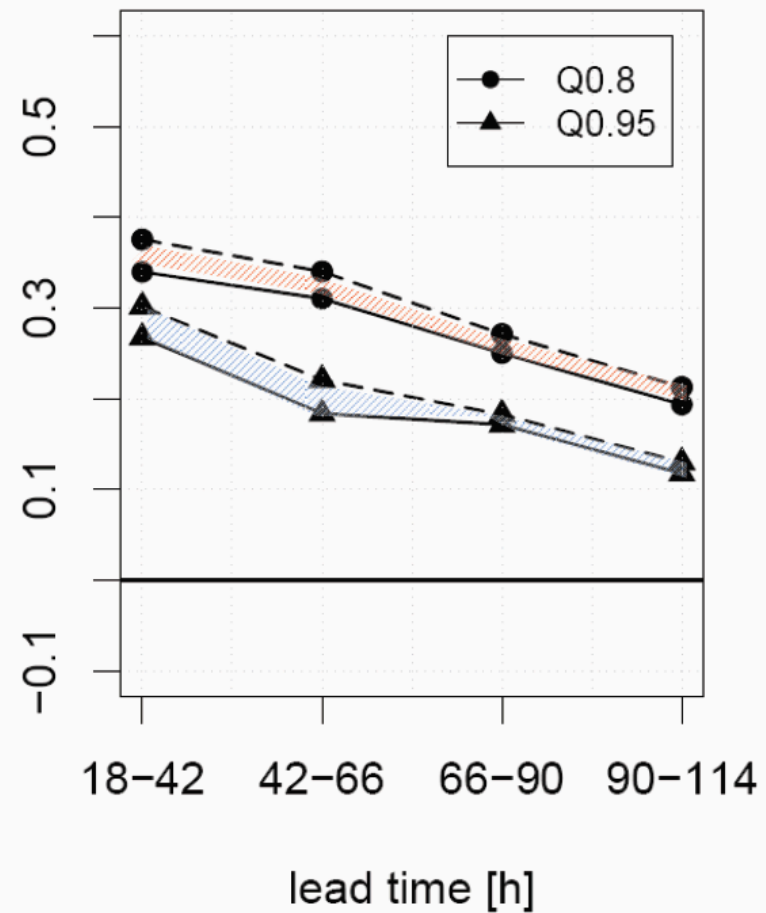
With warm-season QPF, the comparatively coarse resolution and parameterized convection in operational 30-km ensemble system produces a forecast that is clearly inferior to the 4-km, resolved convection ensemble. Statistical downscaling will only work if the model generally has some predictable signal at the larger scales.



winter



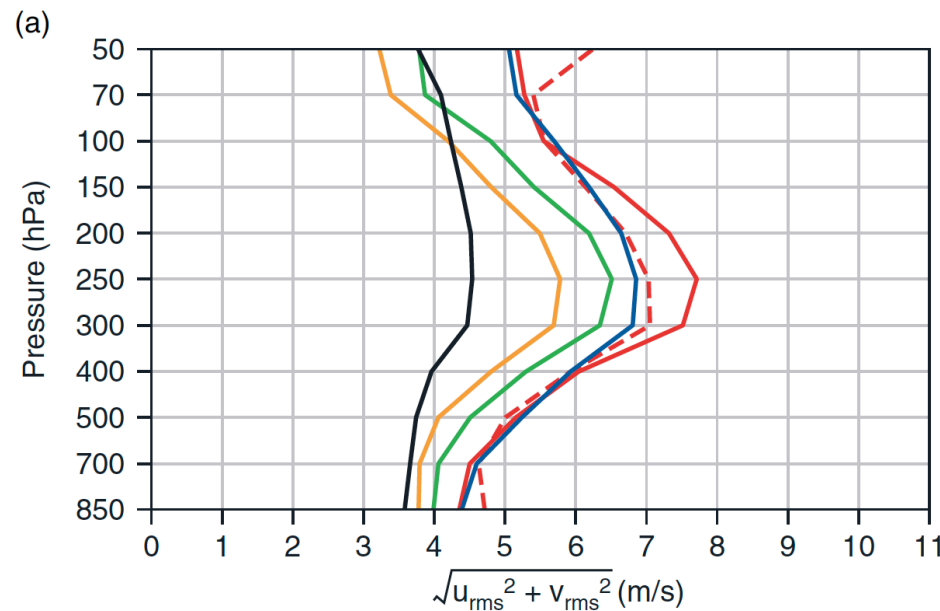
summer



Disadvantage: non-stationary forecast errors in reforecasts?

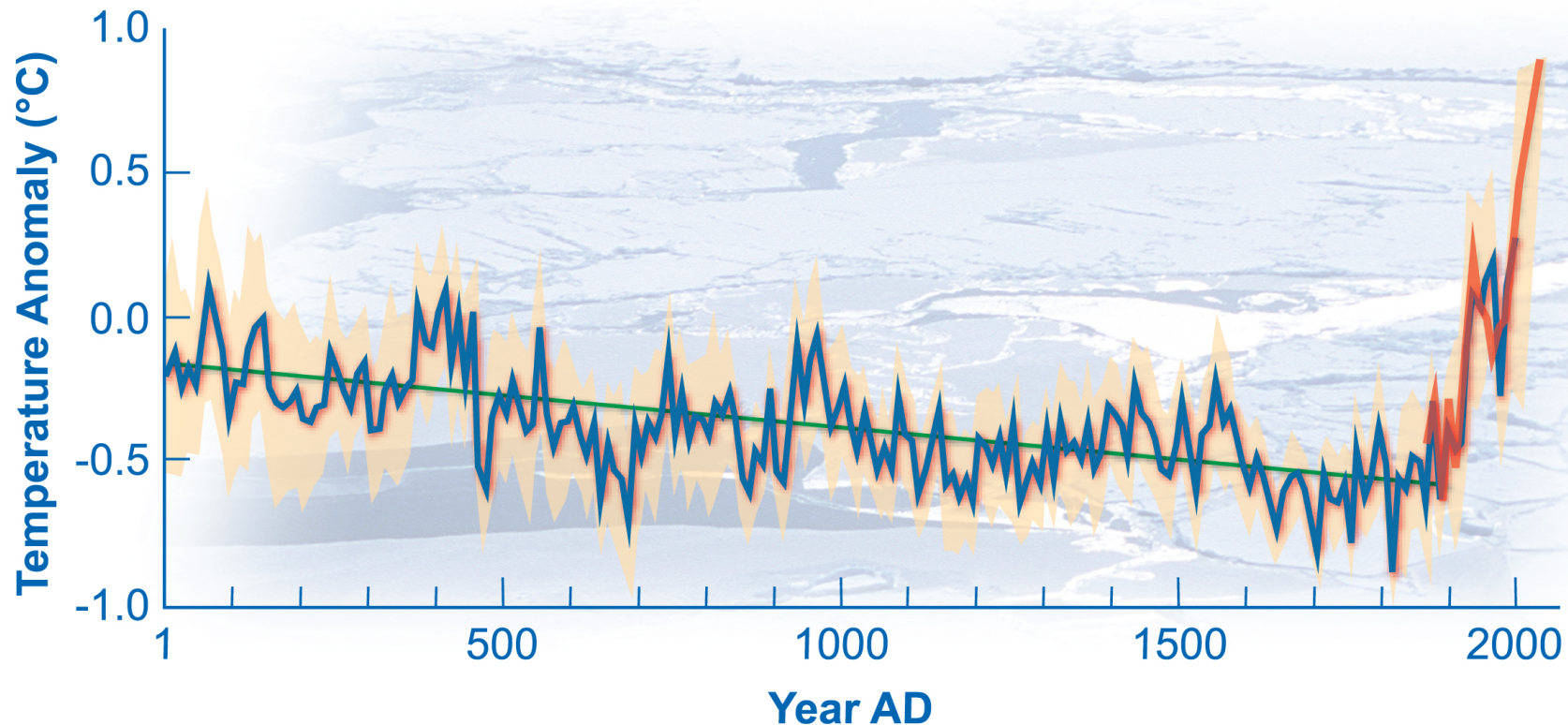
- If real climate or model-error statistics change significantly during reforecast period, decreased accuracy of post-processed estimates.

Short-term forecast fit to radiosondes



— FGGE Main June 1979 - - FGGE Final June 1979 — ERA-15 full year 1979 — ERA-40 June 1979 — ERA-Interim June 1979 — Operations June 2007

Changing climate: today's forecasts warmer than those in training data set?



If forecast today is warmer than any in the reforecast training data set, we'll be “extrapolating the regression” when we apply statistical corrections.

Computational burden of reforecasting

- Real-time ensemble: assume 50 members, 2x daily = 100/day = 700/week
- Minimal reforecast: 5 members, 20 years, 1x weekly = 100/week : **1/7 extra**
- Moderate reforecast: 10 members, 30 years, 1x daily = 2100/week : **3x extra.**
- Full reforecast: 50 members, 30 years, 2x daily = 21000/week : **30x extra!**

Two general approaches to generating a reforecast data set

- Compute once, freeze the model (our approach). Repeat only once every couple years.
 - Offline calculation of reforecasts, we can afford to generate a larger number of them.
- Real-time computation. For whatever model version you're running, make sure you have reforecasts to calibrate (ECMWF approach).

What reforecast data sets are or will be available for NWP?

Available weather reforecast data sets

Producer	# years	# members	frequency	real-time or offline?	resolution	forecast duration
ECMWF EPS	18	5	weekly	real-time	T639 then T319 after 10 days	30 days
NCEP GEFS (1998 version)	32	15	daily	offline	T62	15 days
NCEP GEFS (late 2011 version)	30+	11	daily	offline	T254 then T190 after 8 days	16 days
COSMO-LEPS	30 (1971-2000)	1	daily	offline	~10 km	90 h

2011 GEFS reforecast: design principles

- Reforecasts will be computed with a (smaller-ensemble) version of the GEFS that will be operational in 2011.
- We hope that GEFS will remain in this configuration, or will be changed only slightly, for several years thereafter.
- Once GEFS changes, either NCEP/EMC or ESRL will continue to run the reforecast version until a next-generation reforecast is in place.

Anticipated configuration

- Every 00Z, every day, 1980-current
- 11-member forecast (control + 10 perturbed)
- Lead time to 16 days.
 - T254L42 to day 8
 - T190L42 from days 7.5 to day 16.
- “CFSR” reanalysis initial conditions (GSI, 3D-Var).
- Mimics the expected operational configuration in late 2011.
- Data saved 3-hourly, to 3-day lead, thereafter every 6 hours.

Storage of reforecast data set

- Storing of “important” agreed-upon subset of data \sim 170 TB. Which fields described in a subsequent slide.
 - ESRL / PSD has purchasing \sim 170 TB of storage and server capability for this data set. Cost \sim \$200K for hardware.
 - Will design software to serve this out to you in several manners (http, ftp, OPeNDAP, etc.).
 - Back this up to tape.
- Storing full 00Z reforecasts and initial conditions \sim 800 TB.
 - Useful for LBC’s to run regional reforecasts, or if more fields added to “important” subset.
 - US Department of Energy will archive this for us.

Proposed fields for “fast” archive

- Mean and every member
- “Fast” archive will be on disk, readily accessible
- Mandatory level data:
 - Geopotential height, temperature, u, v, at 1000, 925, 850, 700, 500, 300, 250, 200, 100, 50, and 10 hPa.
 - Specific humidity at 1000, 925, 850, 700, 500, 300, 250, 200
- PV ($\text{K m}^2 \text{ kg}^{-1} \text{ s}^{-1}$) on $\theta = 320\text{K}$ surface.
- Wind components, potential temperature on 2 PVU surface.

Fixed fields to save once

- field capacity
- wilting point
- land-sea mask
- terrain height

Expected single-level fields for “fast” archive

- Surface pressure (Pa)
- Sea-level pressure (Pa)
- Surface (2-m) temperature (K)
- Skin temperature (K)
- Maximum temperature since last storage time (K)
- Minimum temperature since last storage time (K)
- Soil temperature (0-10 cm; K)
- Volumetric soil moisture content (proportion, 0-10 cm) –
- Total accumulated precipitation since beginning of integration (kg/m^2)
- Precipitable water (kg/m^2 , vapor only, no condensate)
- Specific humidity at 2-m AGL (kg/kg ; instantaneous) –
- Water equivalent of accumulated snow depth (kg/m^2) –
- CAPE (J/kg)
- CIN (J/kg)
- Total cloud cover (%)
- 10-m u- and v-wind component (m/s)
- 80-m u- and v-wind component (m/s)
- Sunshine duration (min)
- Snow depth water equivalent (kg/m^2)
- Runoff
- Solid precipitation
- Liquid precipitation
- Vertical velocity (850 hPa)
- Geopotential height of surface
- Wind power ($=\text{windspeed}^3$ at 80 m* density)

Proposed fields for “fast” archive

- Fluxes (W/m^2 ; average since last archive time)
 - sensible heat net flux at surface
 - latent heat net flux at surface
 - downward long-wave radiation flux at surface
 - upward long-wave radiation flux at surface
 - upward short-wave radiation at surface
 - downward short-wave radiation flux at surface
 - upward long-wave radiation at nominal top
 - ground heat flux.

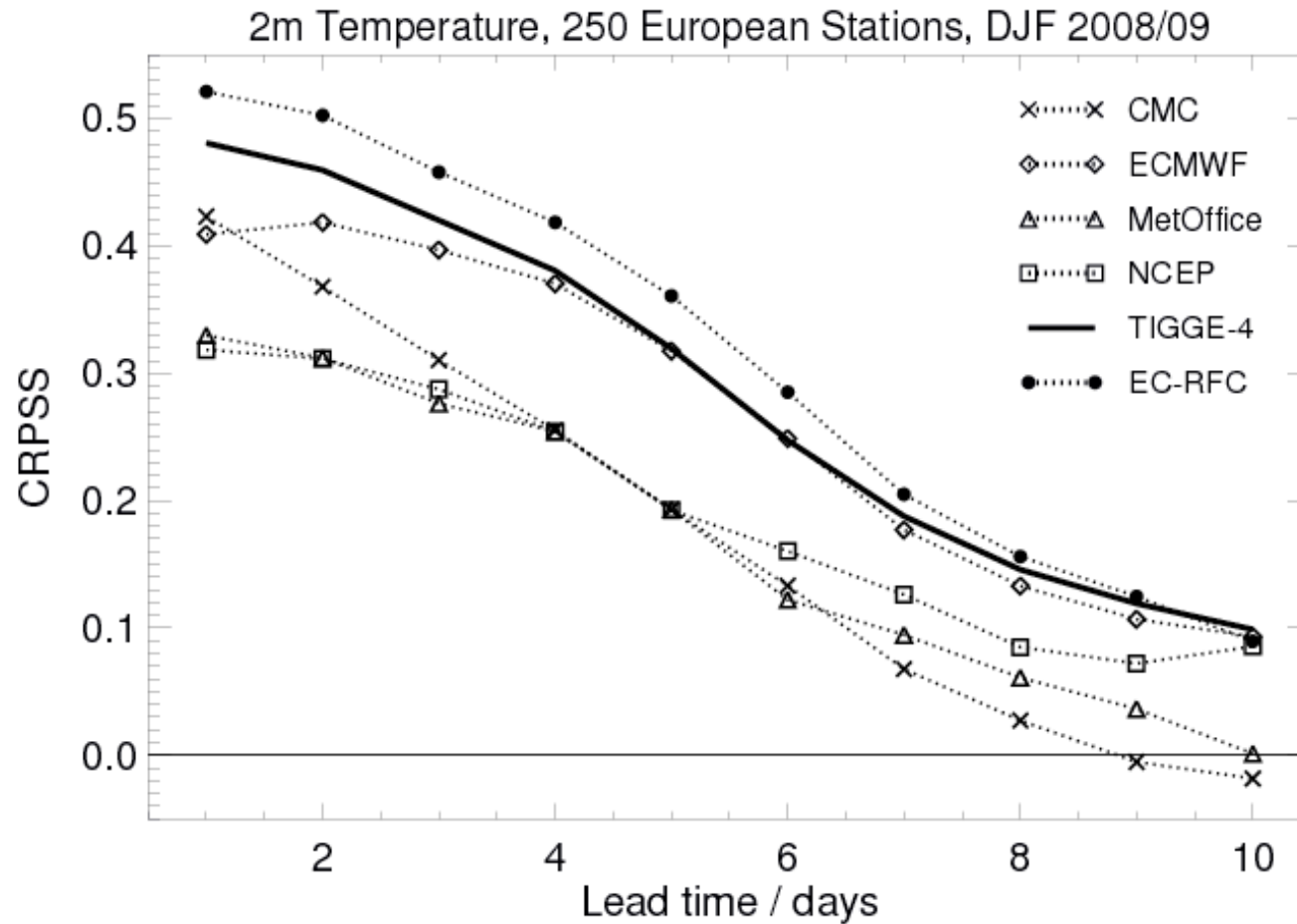
Where we are now

- 170 TB storage in place.
- Control run done.
- Perturbed initial conditions done.
- Are just beginning to compute the perturbed member reforecasts.

Issues in reforecasting

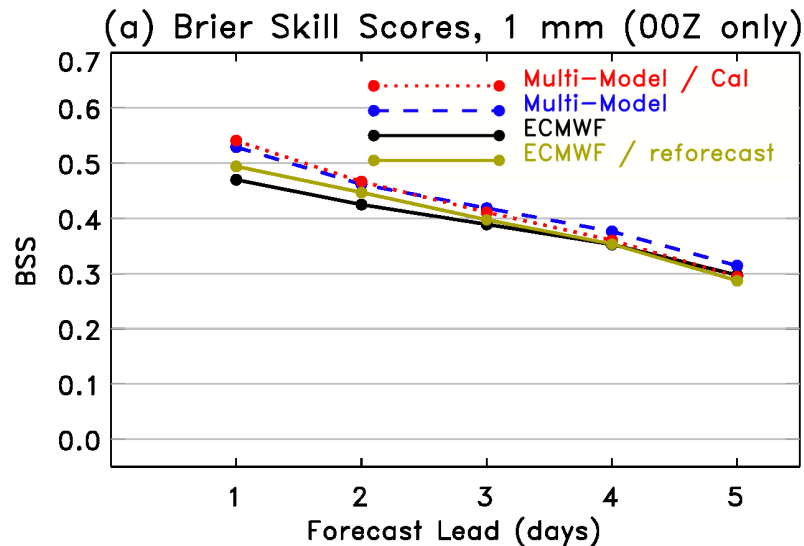
- Relative benefits of statistical vs. dynamical downscaling for particular phenomena.
- What is the appropriate compromise (large reforecast, large computational burden) vs. (small reforecast, small computational burden).
- What are the best statistical post-processing techniques for a particular phenomenon
 - may differ for T_{sfc} , precipitation, hurricane track, severe weather likelihood, etc.
- Examining errors in low-frequency processes in the model (MJO, NAO, etc.).
- Relative benefits of multi-model vs. single-model + reforecast.

Reforecast vs. multi-model, T_{sfc}

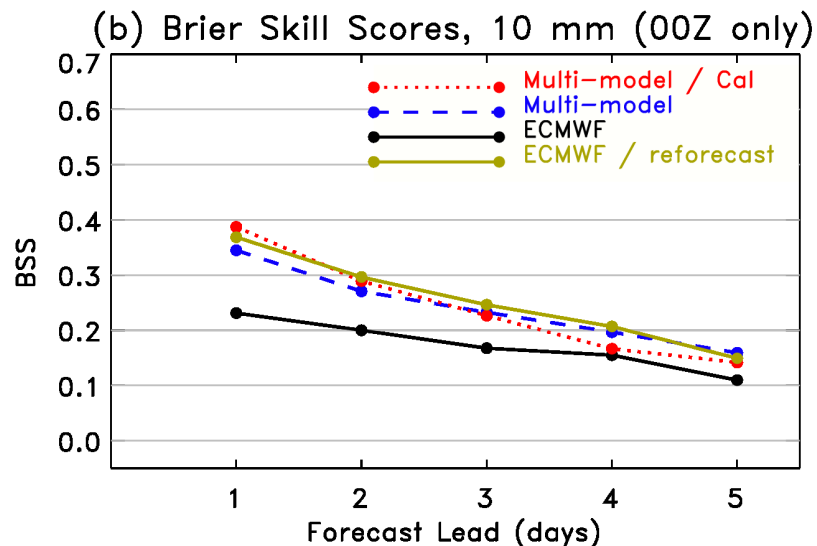


courtesy of Renate Hagedorn, ECMWF & DWD

Reforecast vs. multi-model precipitation over US, Jul-Oct 2010



Now, the following forecasts are plotted: 20-member ECMWF forecasts (black); ECMWF, calibrated via logistic regression using 9 years of ECMWF 4-member weekly reforecasts (green); multi-model (blue) and multi-model, calibrated using the last 30 days of forecasts/analyses.

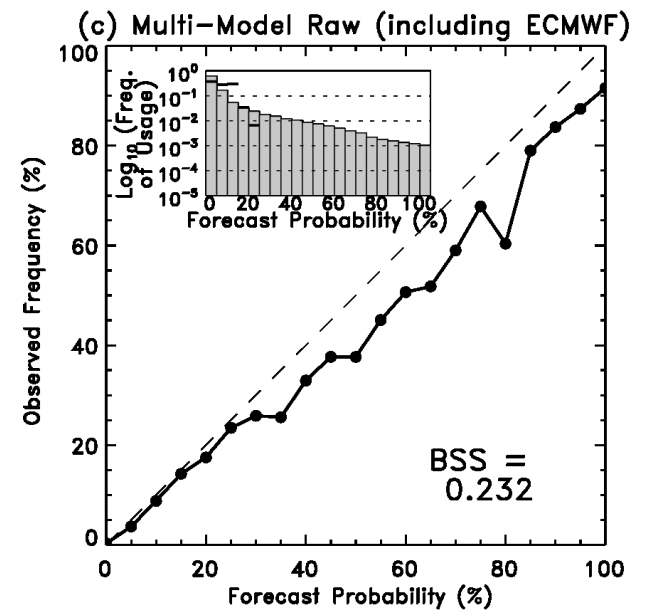
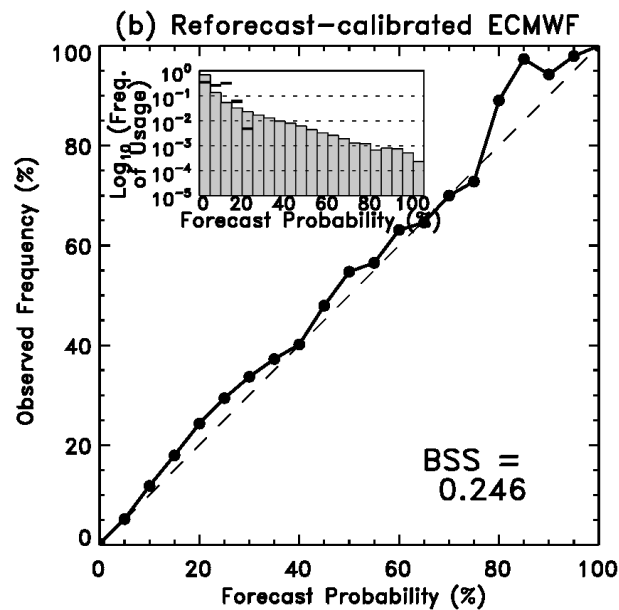
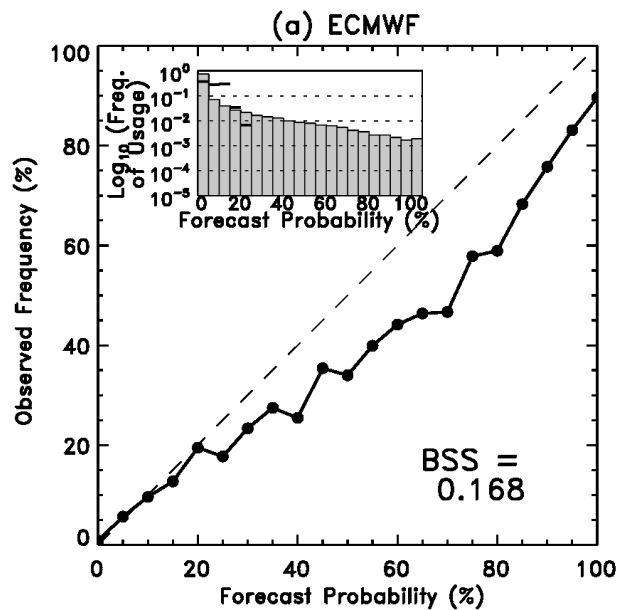


Reforecasts appear to provide most improvement at heavy precipitation thresholds.

Sample reliability diagrams

ECMWF, reforecast-calibrated, multi-model

Reliability, Day +3 10.0mm



(more reliable)

(sharper)

Conclusions

- We believe reforecast data sets can be used to improve forecast guidance for a wide range of phenomena.
- There are still many issues where your research can help us understand the role of reforecasting in the larger weather prediction effort.
- We look forward to further discussions and possible collaborations.

Statistical post-processing, holistic view

- Would like probability distribution ϕ of true state T (or samples thereof) given all available information, including today's ensemble forecast

$$\phi\left(T \mid \mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{z}\right)$$

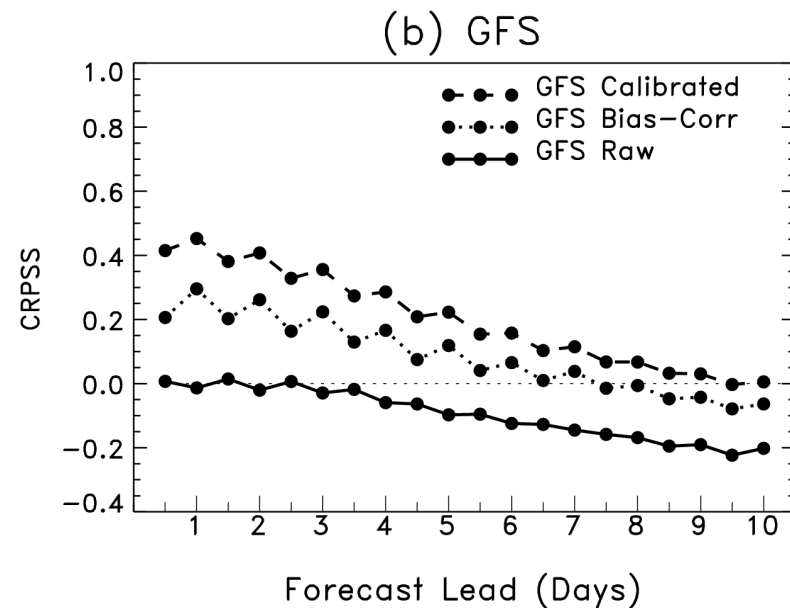
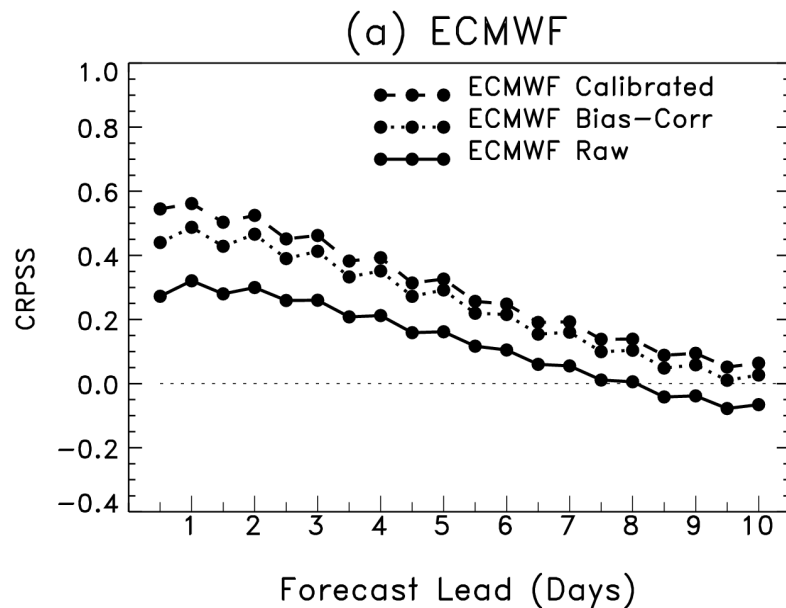
- \mathbf{z} might be past observations and/or forecasts, output from other modeling systems, your intuition, etc.
- ϕ might be a field rather than a scalar

A quick survey of common post-processing techniques

- Simpler methods
 - Gross bias correction
 - Kalman-inspired filters
 - CDF-based bias corrections
 - Linear regression
- Some more complex methods
 - Logistic regression
 - Analog approach
 - Bayesian model averaging (MDL's EKDMOS very similar)
 - Bayesian processor of forecasts
 - Non-homogeneous Gaussian regression
 - Rank histogram-based calibration

Gross bias correction

- Given sample of past forecasts x_1, \dots, x_n and observations y_1, \dots, y_n , gross bias correction is simply $\bar{y} - \bar{x}$



In surface-temperature calibration experiments with NCEP's GFS and ECMWF, simple gross bias correction achieved a large percentage of the improvement that was achieved through more sophisticated, bias+spread correction.

Gross bias correction

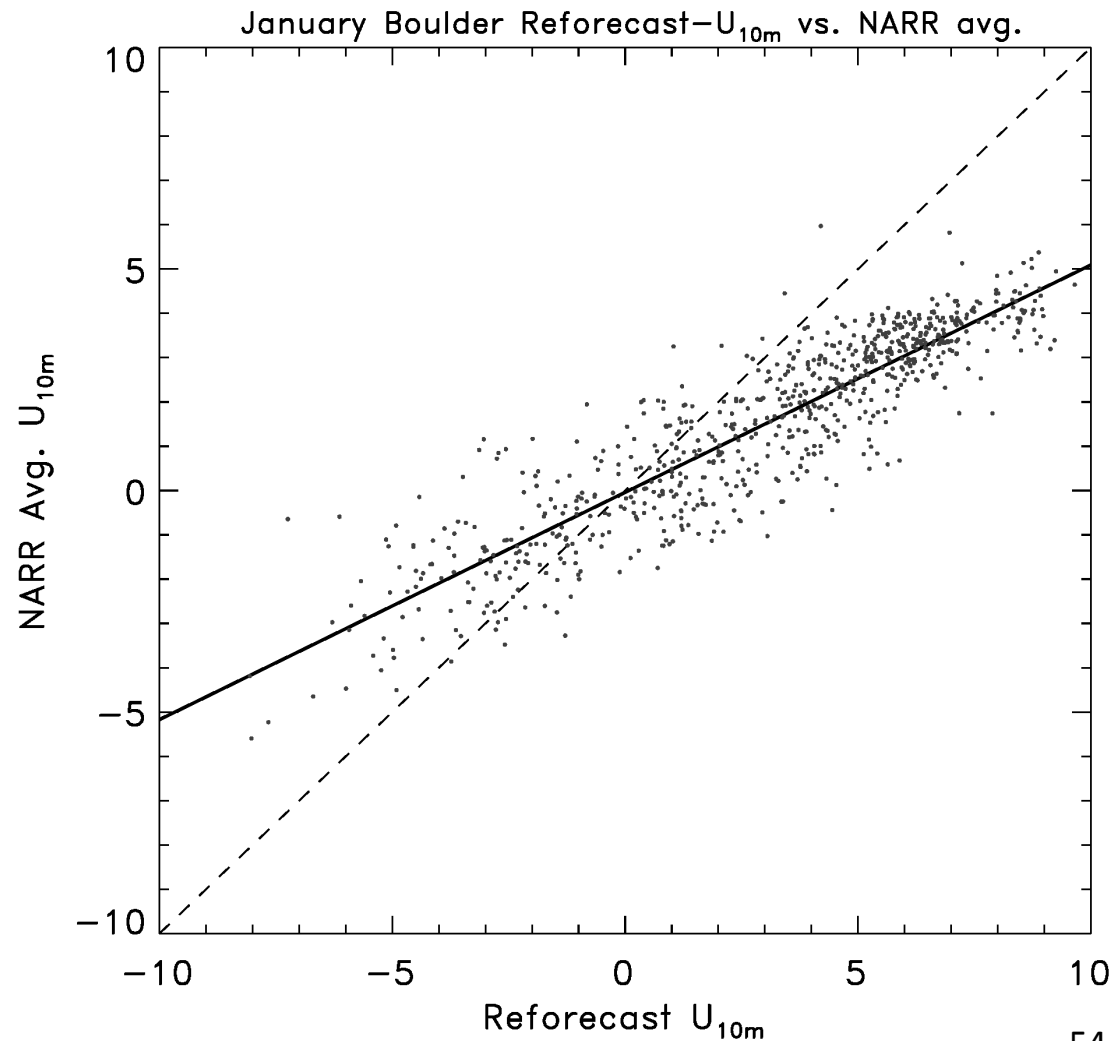
- Effectively, the implied statistical model is the following:

$$Y_i = \beta + X_i + \varepsilon_i$$

- assumes normality of errors; uncorrelated errors, error not *state dependent* (next slide).

State-dependent errors

For this 10-m wind, the bias is **conditional**, depends on the forecast amount. Linear regression (discussed later) a much better choice.



“Kalman-inspired” filter

Today's forecast bias estimate

Yesterday's bias estimate

Yesterday's observed bias

$$\hat{b}_t^f = \hat{b}_{t-1}^f + K_t \left(\varepsilon_t - \hat{b}_{t-1}^f \right)$$

Kalman gain: weighting applied to residual. Larger K_t , more weight to recent data, and vice versa.

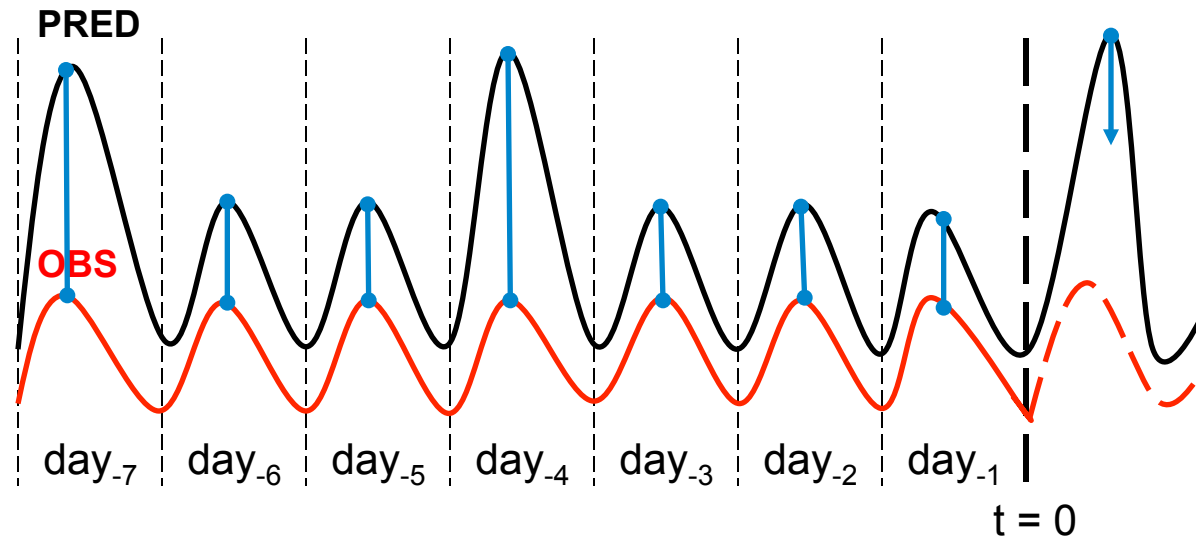
Pro:

- memory in system, amount tunable through K_t
- adaptive

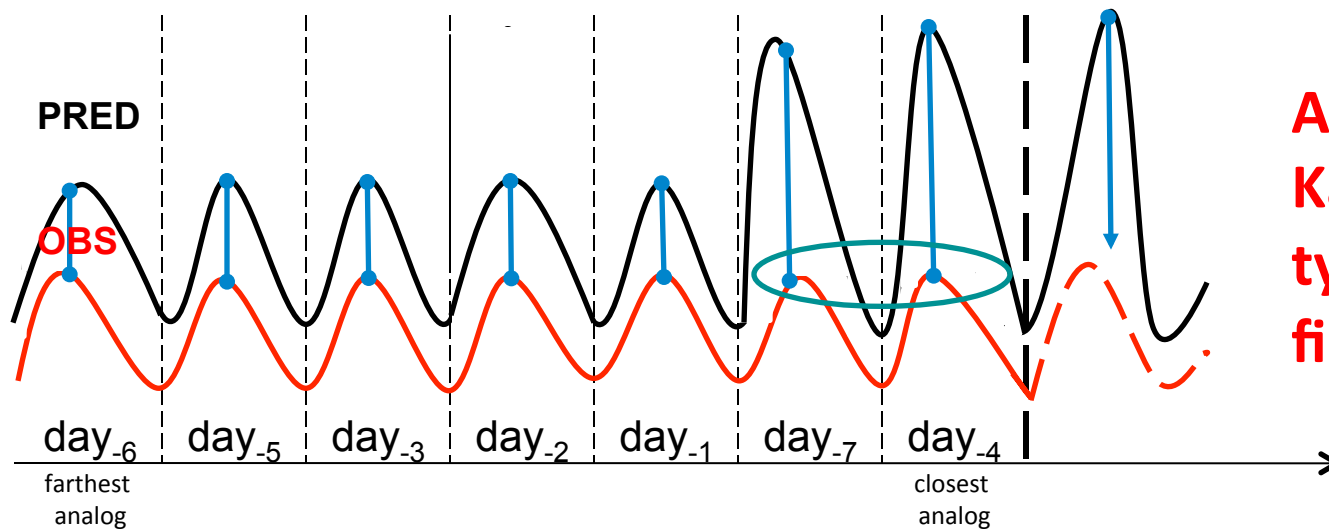
Con:

- assuming there is state-dependent bias, takes time to adapt after regime change and change of state.

An alternative “analog” formulation

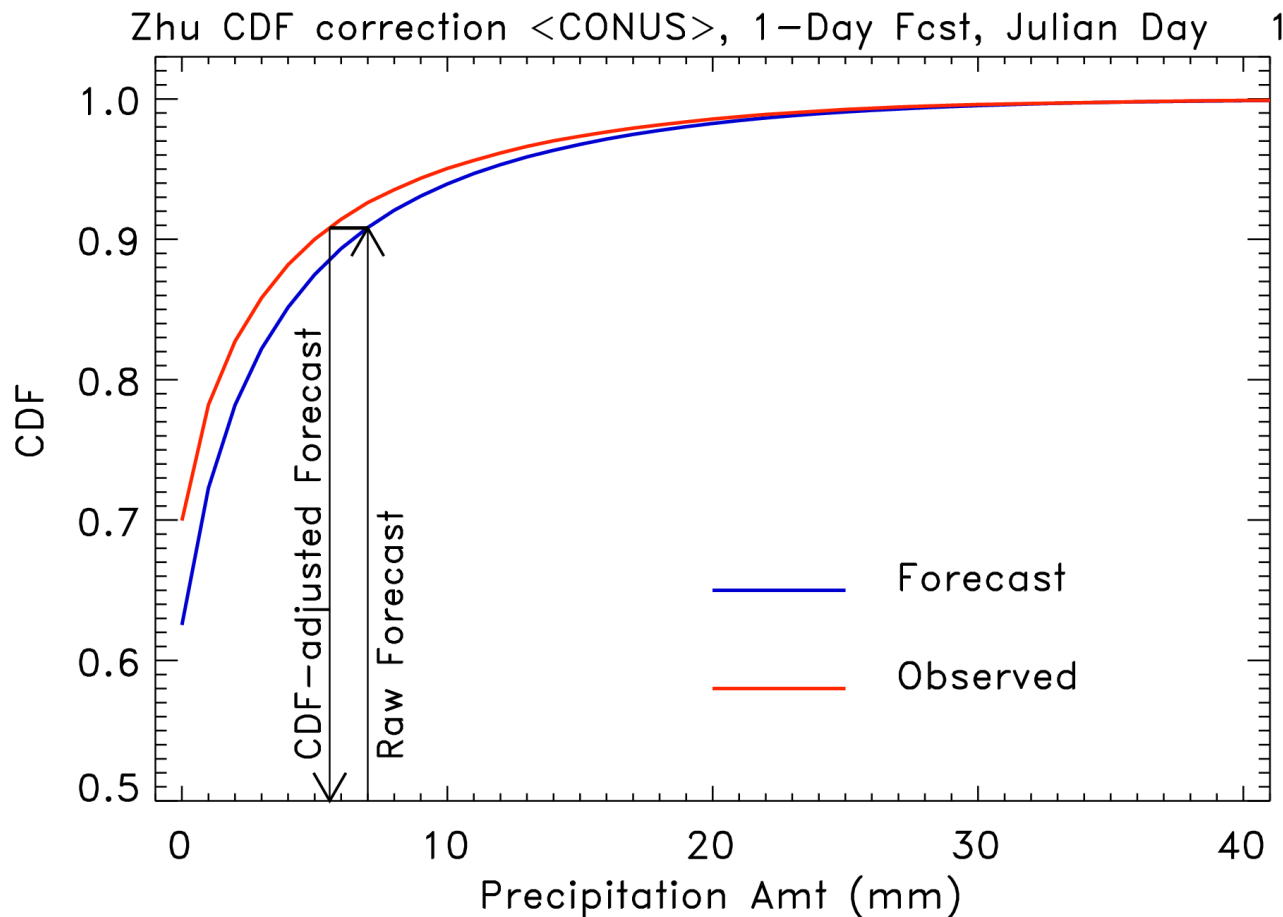


**Standard
Kalman-type
filter**



**Analog
Kalman-
type
filter**

CDF-based bias corrections

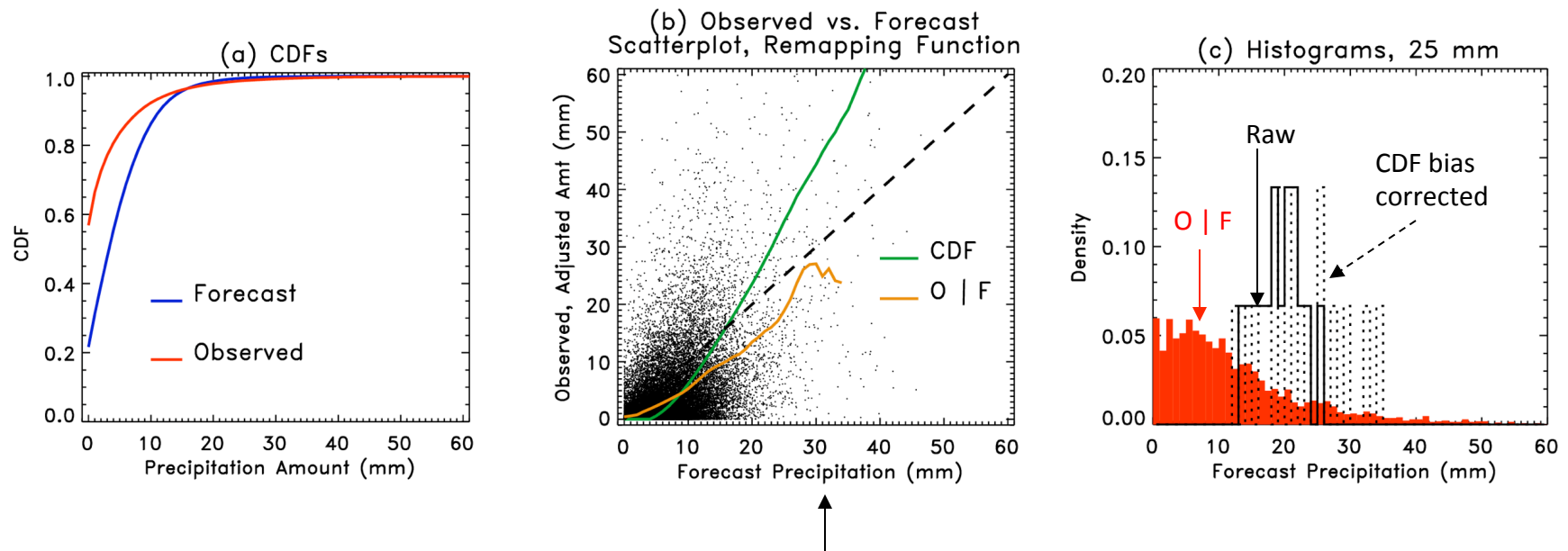


Use difference in CDFs to correct each ensemble member's forecast. In example shown, raw 7-mm forecast corrected to ~5.6 mm forecast.

NOTE: bias only, not spread correction or downscaling.

CDF corrections: example of problem

1-day forecasts in Northern Mississippi (US), mid-August.
Consider a forecast precipitation of 25 mm.

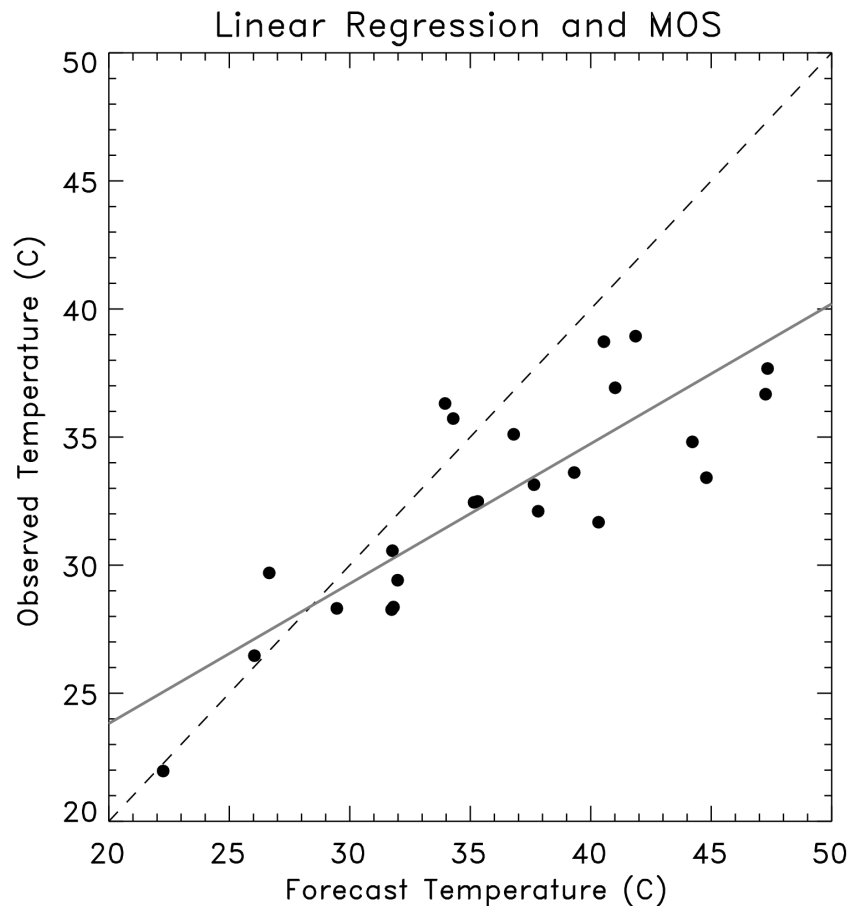


CDF-based corrections at high amounts suggest further increasing precipitation amount forecast. O|F indicates decrease.

At root of problem is assumption that $\text{Corr}(F, O) \approx 1.0$

Linear regression

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_N X_{i,2} + \varepsilon_i$$




Corrects for state-dependent bias; when no predictive skill of forecast, regresses to observed sample climatology.

Diagnostics include statistics on error, so can infer (largely non-state dependent) pdf.

When is linear regression approach useful?

- Some assumptions:
 - Normality of errors.
 - Linear relation between predictors and predictand.
 - Homoscedasticity, error variance doesn't depend on state x .
 - Errors are uncorrelated between samples.

When is linear regression approach useful?

- Some assumptions:
 - Normality of errors.
 - Linear relation between predictors and predictand.
 - Homoscedasticity, error variance doesn't depend on state x .  Well, there goes using this for precipitation!
 - Errors are uncorrelated between samples.

When is linear regression approach useful?

- Some assumptions:
 - Normality of errors.
 - Linear relation between predictors and predictand.
 - Homoscedasticity, error variance doesn't depend on state x .
 - Errors are uncorrelated between samples.



Problematic for weather, if samples every day.
“Serial correlation,” smaller “effective sample size.”
But can deal with this problem.

Linear regression – big assumption!

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_N X_{i,2} + \varepsilon_i$$

↑ unknown, with error

↑ known, assumed no error

↑

In our practice, the Y's typically have some small error (obs) and the X's have larger error (forecast model state).

Practically, the method works well enough to gloss over what error ε_i represents, but there is a whole branch of statistics (regression with “errors in variables”) that deals with this more formally. This incorrect assumption applies to most of the rest of the methods discussed, too.

Model Output Statistics (“MOS”)

most elements based on multiple linear regression

KBID	GFS	MOS	GUIDANCE												2/16/2005	1800	UTC																																																																	
DT	/FEB	17													/FEB	18											/FEB	19																																																						
HR	00	03	06	09	12	15	18	21	00	03	06	09	12	15	18	21	00	03	06	12	18																																																													
N/X													32													40							25							35							19																																			
TMP	42	39	36	33	32	36	38	37	35	33	30	28	27	30	32	31	28	25	23	19	27																																																													
DPT	34	29	26	22	19	18	17	17	17	17	17	15	14	13	11	8	7	6	5	2	4																																																													
CLD	OV	FW	CL	CL	SC	BK	BK	BK	BK	BK	BK	SC	BK	BK	BK	BK	FW	CL	CL	CL																																																														
WDR	26	30	32	32	32	31	29	28	30	32	31	31	31	31	30	29	31	32	33	33	27																																																													
WSP	12	12	12	11	08	08	09	08	09	09	10	10	10	12	13	13	15	16	15	09	08																																																													
P06													17													0							0							4							0							10							6							8							0							0
P12													17													0							10							17							8																																			
Q06													0													0							0							0							0							0																												
Q12													0													0							0							0							0																																			
T06													0/ 2													0/ 0							1/ 0							1/ 2							0/ 1							0/ 1							1/ 0							0/ 1							0/ 0							0/ 0
T12													1/ 0													1/ 2							1/ 1							0/ 1							0/ 0																																			
POZ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0																																																													
POS	13	47	70	84	91	100	96	100	100	100	100	100	92	100	98	100	100	100	100	94	92	100	100																																																											
TYP	R	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S																																																													
SNW													0													0							0																																																	
CIG	7	8	8	8	8	8	8	8	8	7	7	7	8	7	7	7	8	8	8	8	8																																																													
VIS	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7																																																													
OBV	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N																																																													

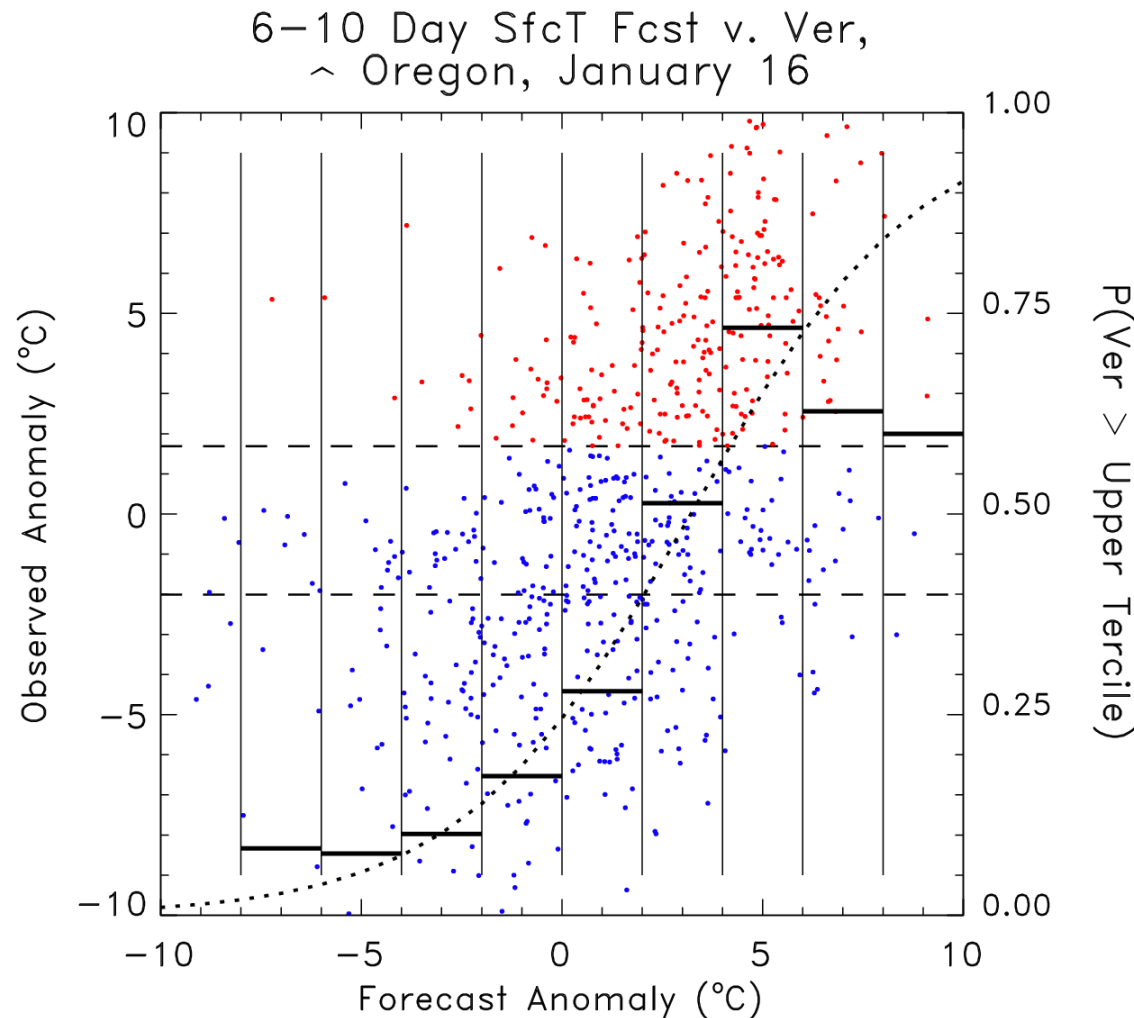
US: Statistical corrections to operational US NWS models, some fixed (NGM), some not (Eta, GFS). Refs: <http://www.nws.noaa.gov/mdl/synop/index.htm>, Carter et al., *WAF*, **4**, p 401, Glahn and Lowry, *JAM*, **11**, p 1580. **Canadian** models discussed in Wilson and Vallee, *WAF*, **17**, p. 206, and *WAF*, **18**, p 288. **Britain:** Met Office uses “updateable MOS” much like perfect prog.

Logistic regression

- Useful for making probabilistic forecasts for some binary event, e.g, precip above threshold.
- For each grid point (or station) let x = continuous predictor data (ens. mean forecast value), y = binary predictand data (1.0 if predicted event happened, 0.0 if not).
- Problem: Compute $P(y = 1.0 \mid x)$ as a continuous function of x .
- Logistic Regression:

$$P = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)}$$

Logistic regression using a long data set of observed and forecast anomalies

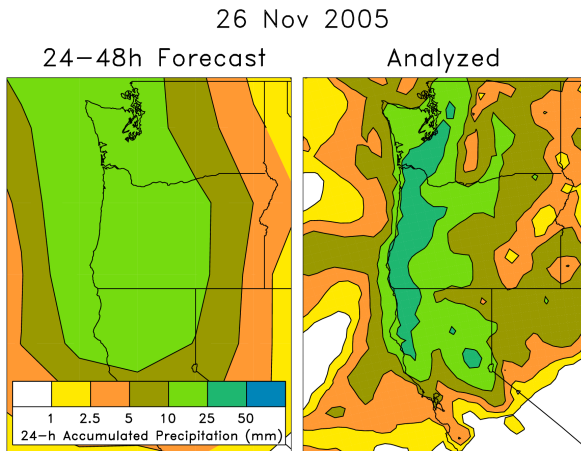


Seeking to predict probability of warmer than normal conditions (upper tercile of observed). Using reforecasts (a later talk), we have 23 years of data. Let's use old data in a 31-day window around the date of interest to make statistical corrections.

Logistic regression drawbacks

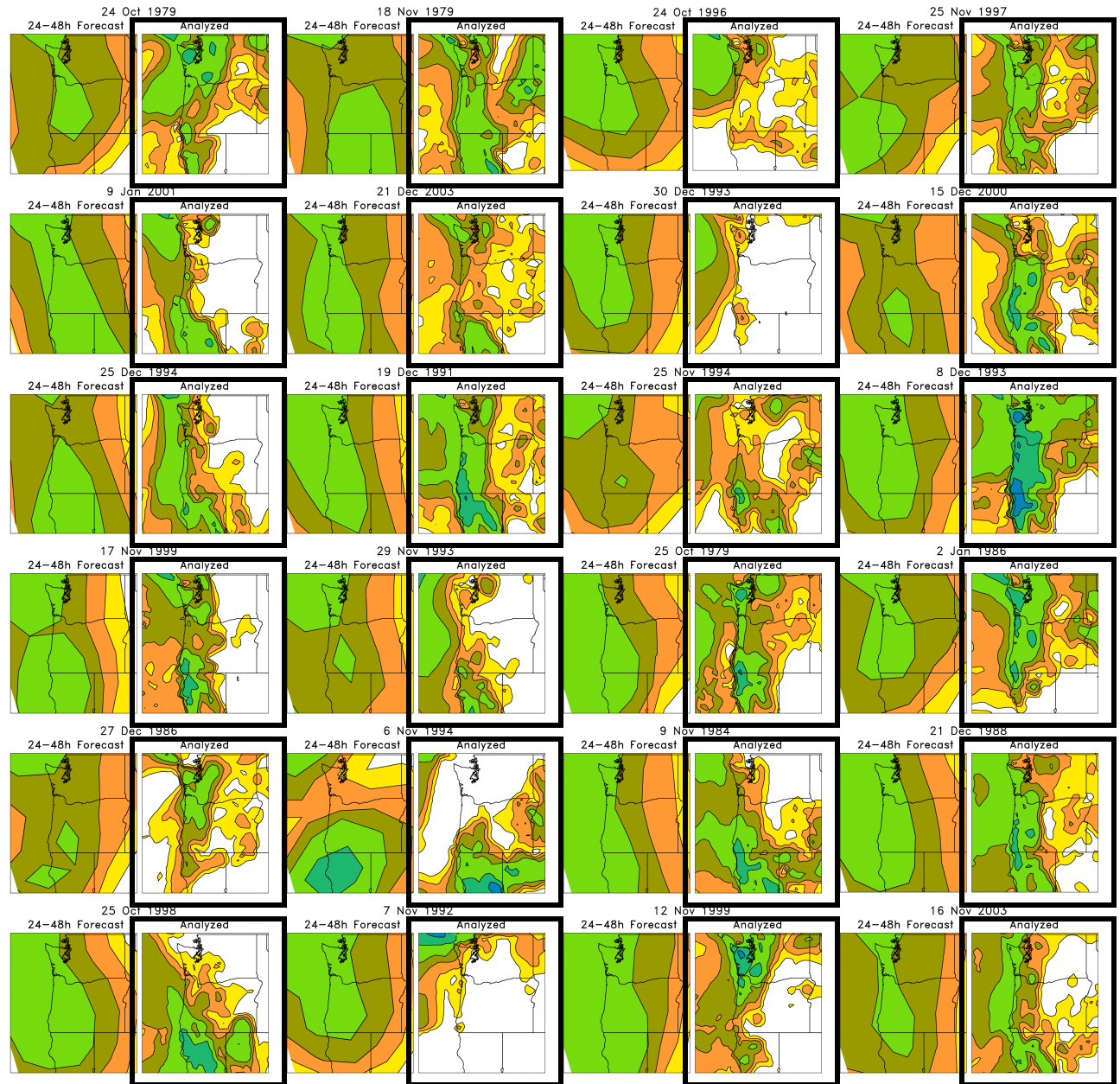
- Doesn't generate full pdf (though see Wilks, 2009, *Met Apps*, p. 361).
- With ensembles, what do you use as predictors? Ens. mean? Spread? Every member?
- Iterative technique, can be slower.
- Better have training set with distribution of 1's and 0's, otherwise software will croak.

Analog technique using reforecasts

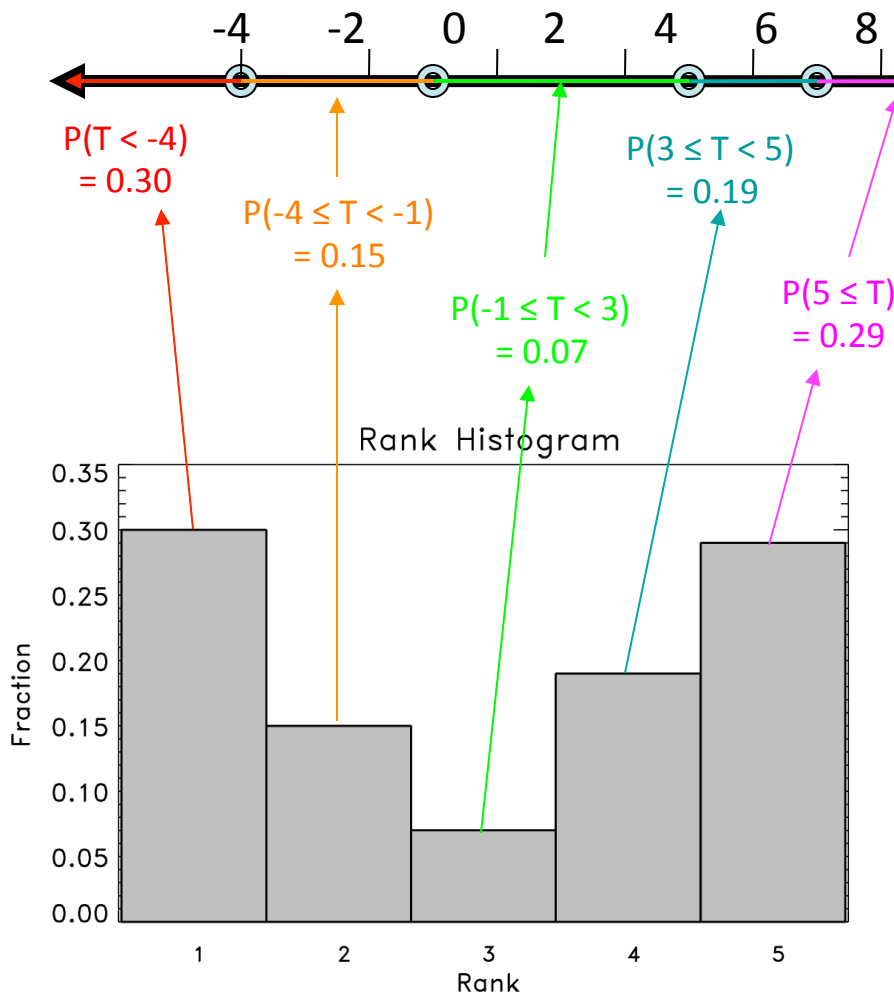


On the left are old forecasts similar to today's ensemble-mean forecast. The data on the right, the analyzed precipitation conditional upon the forecast, can be used to statistically adjust and downscale the forecast.

Analog approaches like this may be particularly useful for hydrologic ensemble applications, where an ensemble of realizations is needed.



Rank histogram technique for ensemble calibration



NCEP MRF precipitation forecasts,
from Eckel and Walters, 1998

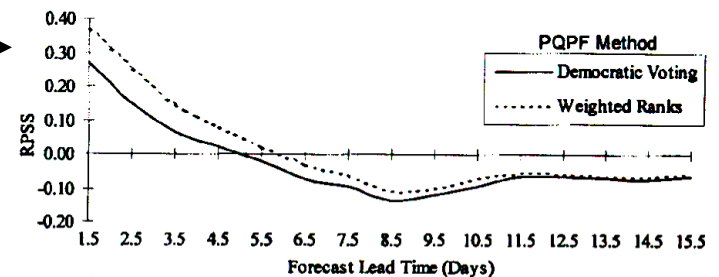


FIG. 10. Ranked probability skill score (RPSS) results for all forecast lead times.

Advantages: Demonstrated skill gain

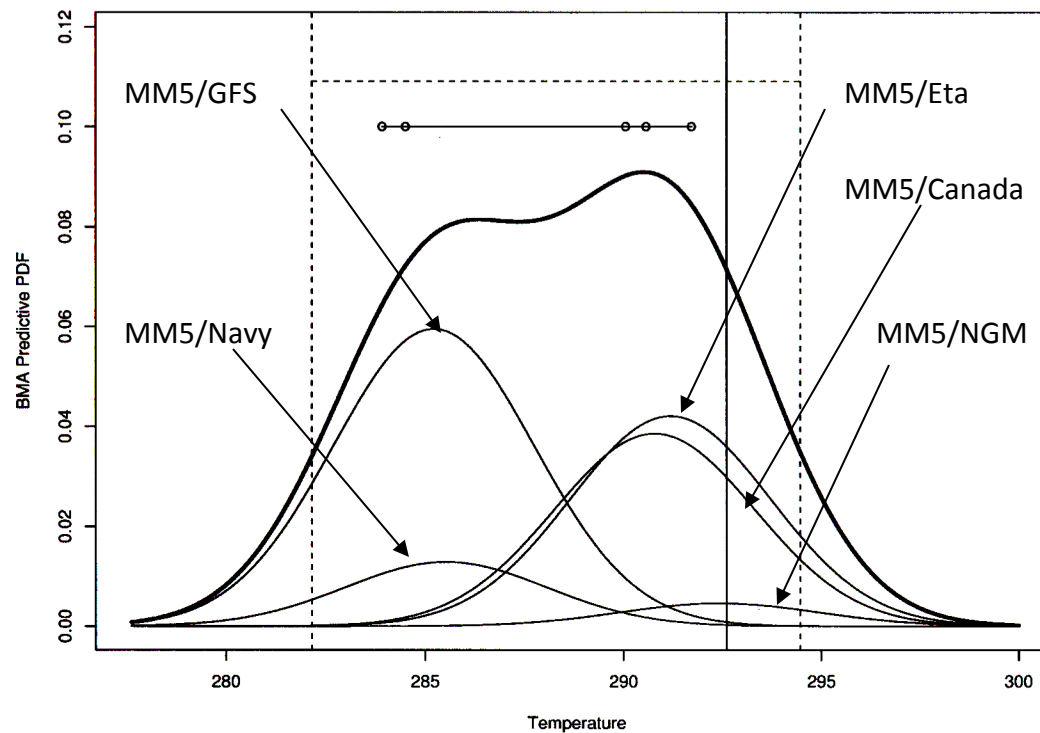
Disadvantages:

- (1) Odd pdfs, especially when two ensemble members close in value.
- (2) Sensitive to shape of rank histogram, and shape of histogram may vary with aspects like precip amount --> sample size issues.
- (3) Fitted parametric distributions as skillful

Bayesian model averaging (BMA)

$$p(y \mid f_1, \dots, f_K) = \sum_{k=1}^K w_k g_k(y \mid f_k) \quad \leftarrow$$

Weighted sum of kernels centered around individual, **bias-corrected** forecasts.



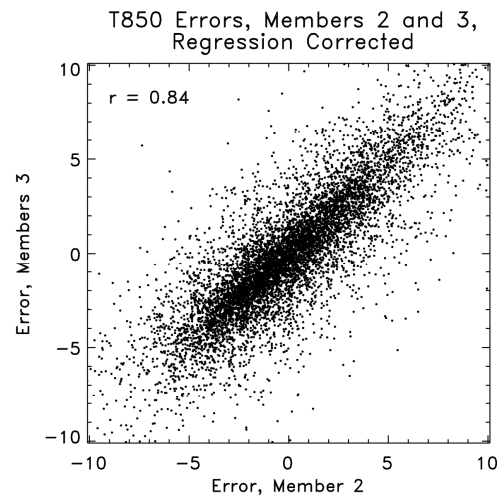
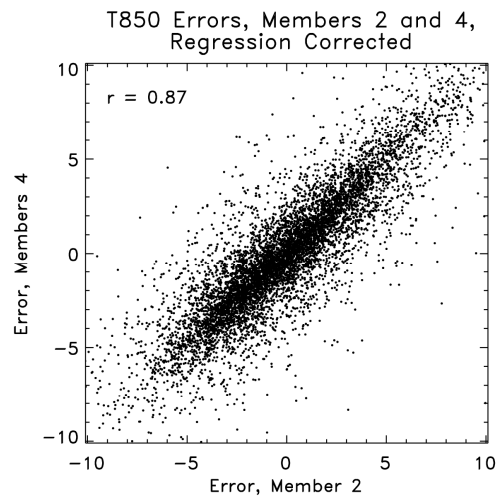
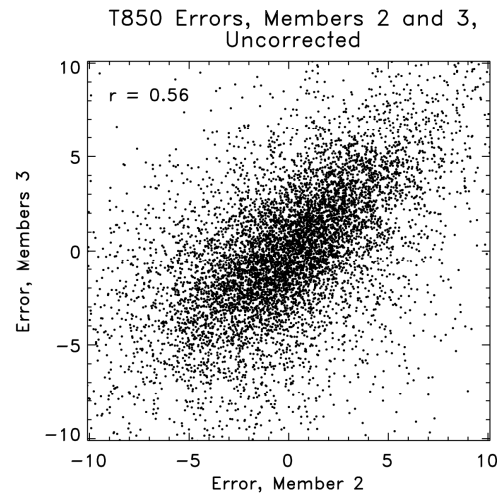
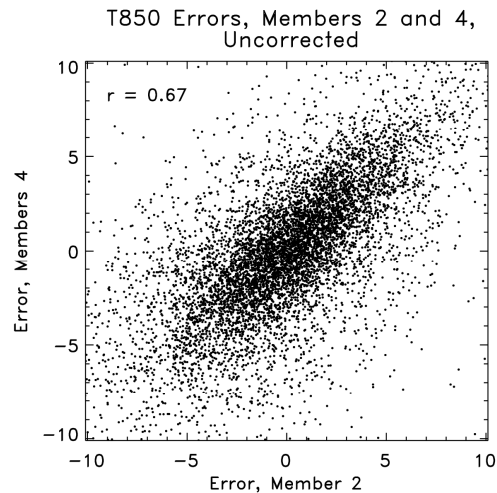
Advantages: Theoretically appealing. No parameterized distribution assumed, weights applied proportional to their independent information (in concept).

Disadvantages: When trained with small sample, **BMA radically de-weighted some members due to “overfitting”** See Hamill, *MWR*, Dec. 2007.

Figure 3: BMA predictive PDF (thick curve) and its five components (thin curves) for the 48-hour surface temperature forecast at Packwood, Wash., initialized at 0000 UTC on June 12, 2000. Also shown are the ensemble member forecasts and range (solid horizontal line and bullets), the BMA 90% prediction interval (dotted lines), and the verifying observation (solid vertical line).

Why BMA's unequal weights?

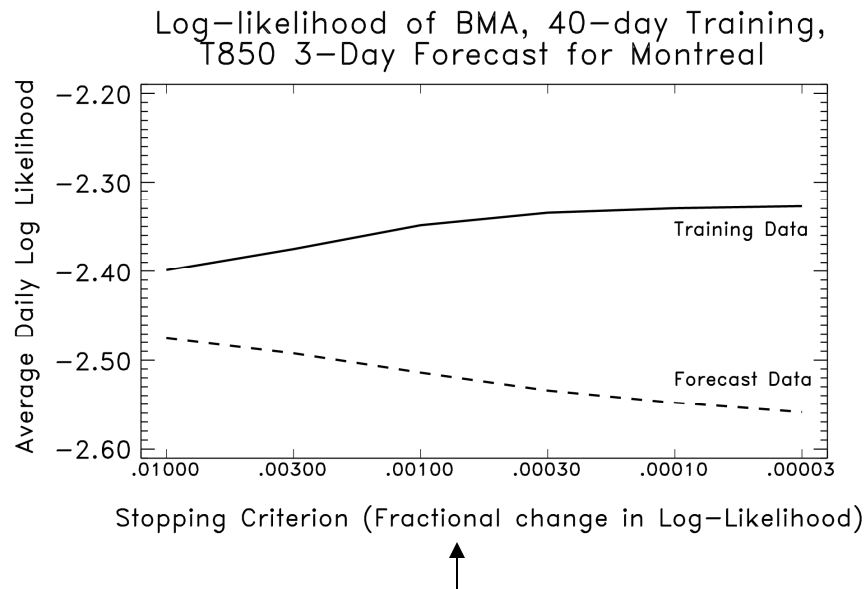
(1) regression correction accentuates error correlations.



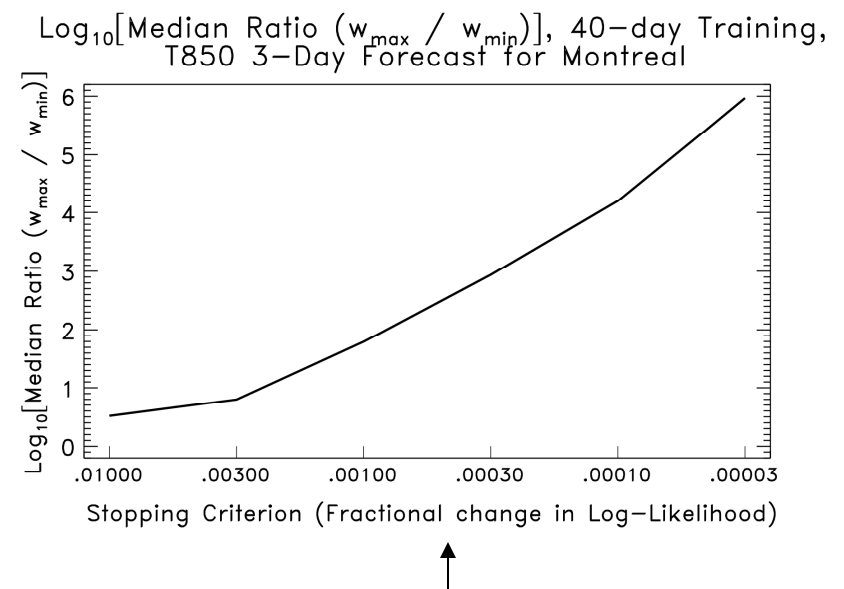
Why BMA's unequal weights?

(2) E-M overfits with little training data

An “estimation-minimization” (E-M) algorithm is used to determine the weights applied to ensemble members. If two forecasts have highly co-linear errors, E-M will weight one very highly, the other very little.

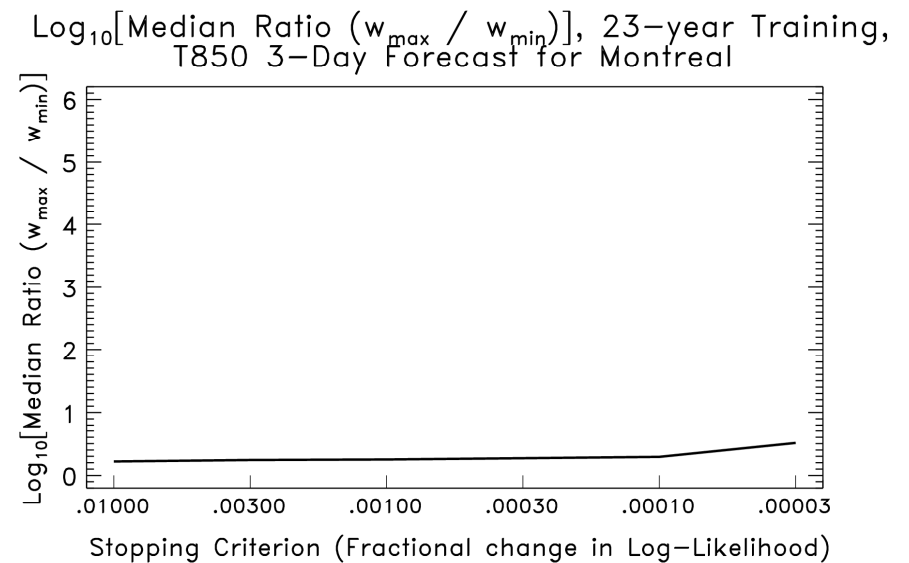
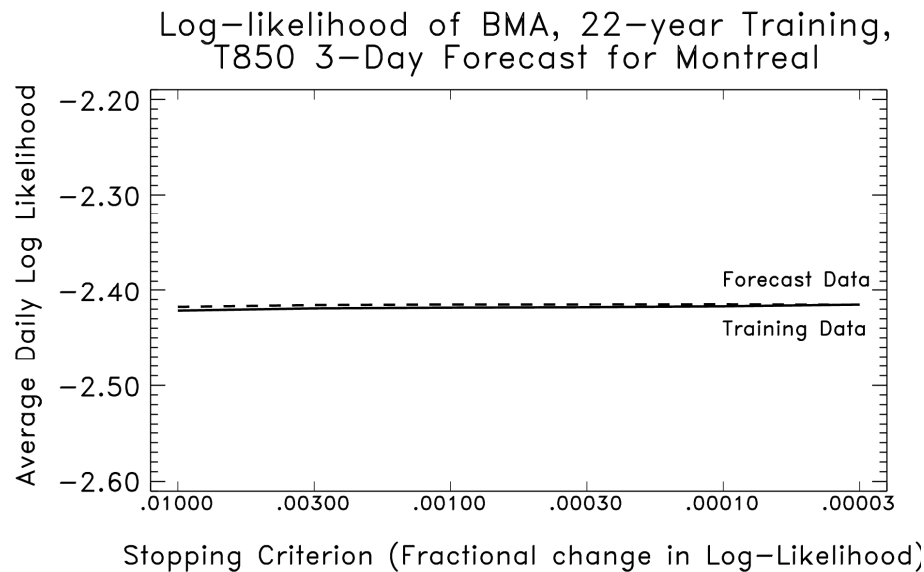


E-M is an iterative technique, and we can measure the accuracy of the fit to the data through the log-likelihood. Something odd happens here; as the E-M convergence criteria is tightened, the fit of the algorithm to independent data gets worse.



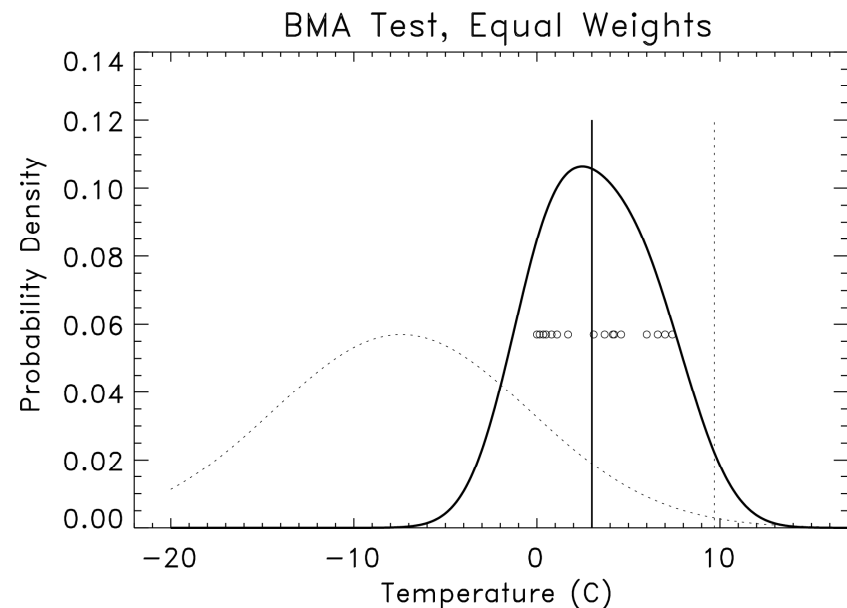
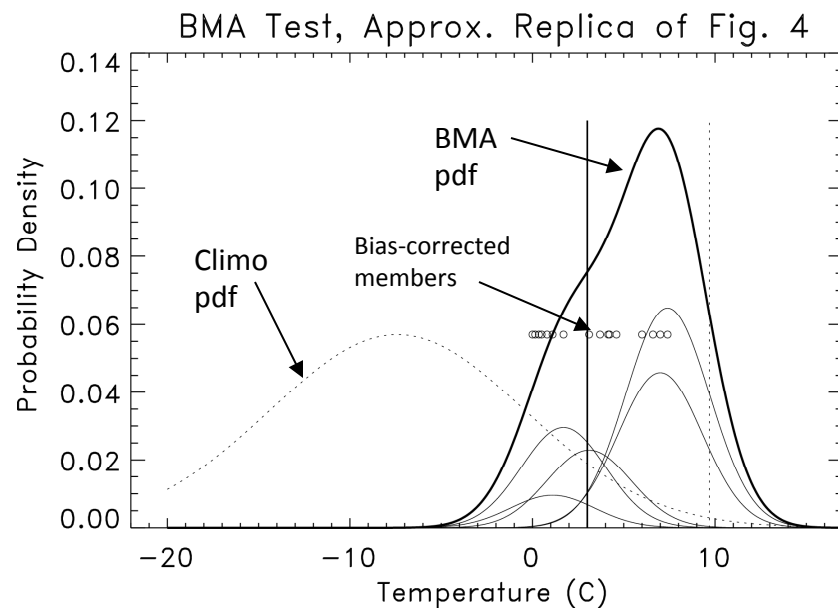
This plots the ratio of the weights of the highest-weighted member to the lowest-weighted member. As the convergence criteria is tightened, the method increasingly weights a few select members and de-weights others.

(BMA overfitting not a problem with 2+ decades training data)



With reforecast data set, we can train with a very large amount of data. When we do so, the weights applied to individual members are much more equal. This indicates that the unequal weighting previously is incorrect.

BMA's problem: an example



Here's a test of BMA in the winter season for a grid point near Montreal. BMA ends up highly weighting the warmest members (inappropriately so), thus producing a very high probability of a warm forecast.

“Bayesian Processor of Forecasts”

- Two key ideas:
 - (1) Bayes’ Rule: leverage prior non-NWP information, whether from climatology, persistence, whatever. Update with NWP information
$$\phi(o|f) \propto \phi(f|o)g(o)$$
 - (2) If data non-normally distributed, transform data to space where normally distributed before performing regression analysis.

Non-homogeneous Gaussian regression

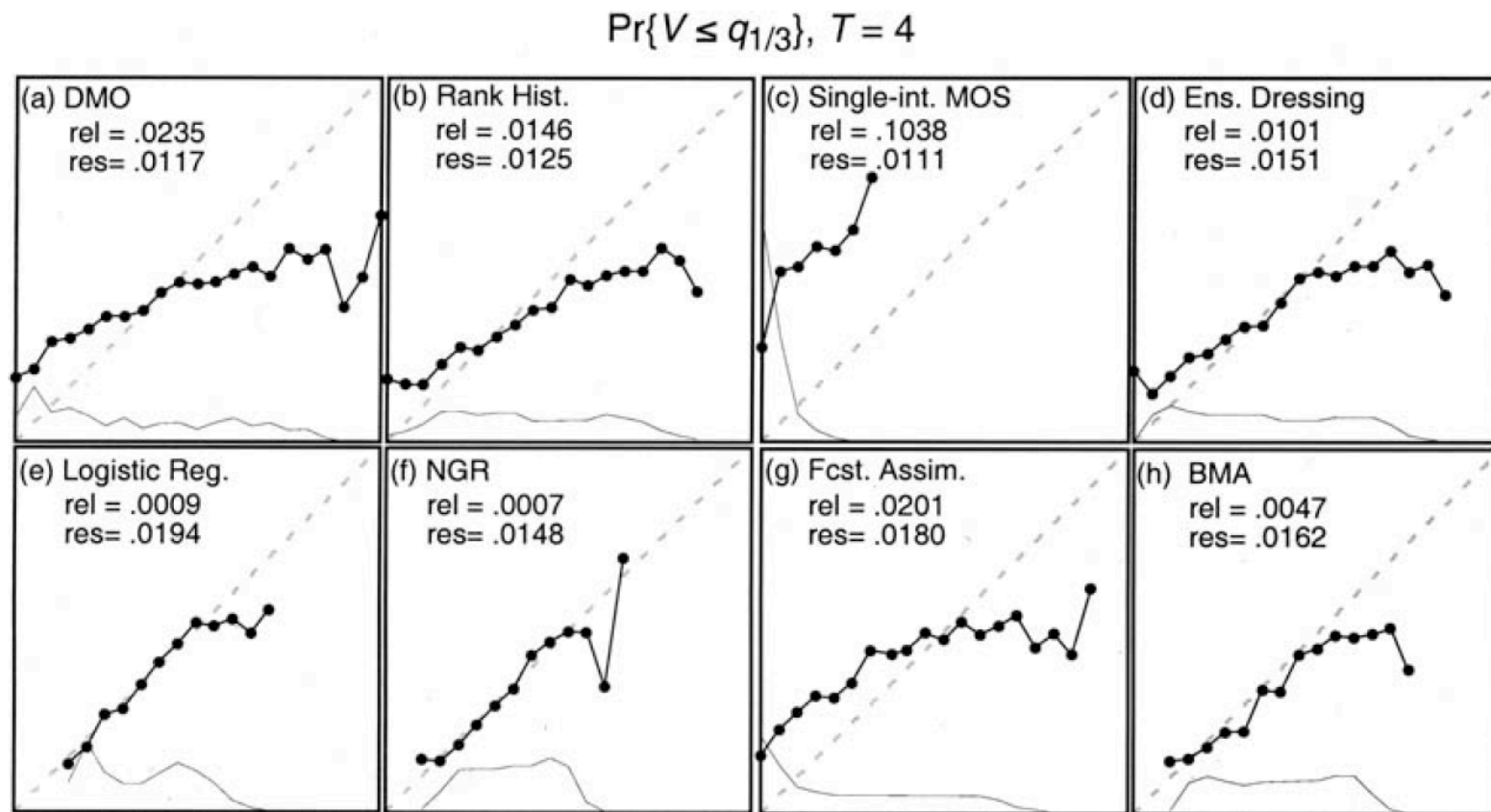
- **Reference:** Gneiting et al., *MWR*, **133**, p. 1098
- **Predictors:** ensemble mean and ensemble spread
- **Output:** mean, spread of calibrated Gaussian distribution

$$f^{CAL}(\bar{\mathbf{x}}, \sigma) \sim N(a + b\bar{\mathbf{x}}, c + d\sigma)$$

- **Advantage:** leverages possible spread/skill relationship appropriately. Large spread/skill relationship, $c \approx 0.0$, $d \approx 1.0$. Small, $d \approx 0.0$
- **Disadvantage:** iterative method, slow...no reason to bother (relative to using simple linear regression) if there's little or no spread/skill relationship.

Is there a “best” calibration technique?

Using Lorenz '96 toy model, direct model output (DMO), rank histogram technique, MOS applied to each member, dressing, logistic regression, non-homogeneous Gaussian regression (NGR), “forecast assimilation”, and Bayesian model averaging (with perturbed members assigned equal weights) were compared. Comparisons generally favored logistic regression and NGR, though differences were not dramatic, and results may not generalize to other forecast problems such as ones with non-Gaussian errors.



77

Figure 8. As Figure 5, for $\Pr\{V \leq q_{1/3}\}$ at lead time $T = 4$.